Interviewer and Respondent Interaction in Survey Interviews

Ongena, Y.P.

2005

Interviewer and Respondent Interaction in Survey Interviews

Thesis committee:

dr. F.G. Conrad (University of Michigan)

dr. B.C. Holleman (Universiteit Utrecht)

Prof.dr. G. Loosveldt (Katholieke Universiteit Leuven)

Prof.dr.W.E. Saris (Universiteit van Amsterdam)

Prof.dr. J. Van der Zouwen (Vrije Universiteit Amsterdam)

ISBN 90-9020070-3

Cover design: S. van der Ploeg, Room for ID's, Nieuwegein

Photography: Hollandse Hoogte

The Netherlands Organization of Scientific Research (NWO) is gratefully acknowledged for funding this project (no. 510-10-027)

VRIJE UNIVERSITEIT

Interviewer and Respondent Interaction in Survey Interviews

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. T. Sminia,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Sociale Wetenschappen
op dinsdag 6 december 2005 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Yfke Pieternel Ongena
geboren te Groningen

promotor: prof.dr. W. Dijkstra
copromotor: dr. A.R. Draisma

Contents

Acknowledgements

This dissertation is the result of a research project that I have worked on with much pleasure for the past five years. I would not have been able to complete this project without the help of several people who gave generously of their time and knowledge, or in any other way supported me.

In the first place, I would like to thank my supervisors Wil Dijkstra and Stasja Draisma. This dissertation greatly benefited from their ideas, comments and advice. Furthermore, their guidance and trust were valuable and highly motivating to me. I appreciate this support very much.

I would also like to thank Ineke Nagel for her suggestions that have improved the presentation of the multilevel analysis, and all other colleagues at the department of social research methodology for comments they gave during colloquia. Many colleagues at the department, especially my fellow Ph.D students, made my stay at the Vrije Universiteit a very pleasant one.

I would not have been able to conduct this research without the data that was available to me. I thank Edit Smit for her kind permission to use the survey data described in chapter 5. The data of the Dutch ESS pilot (described in chapter 6) was made available by Willem Saris.

For the data of the survey described in chapter 7, I owe thanks to my research assistant, Sanne van Krimpen, and to other assistants who took care of transcriptions and/or coding: Eline Rinsema, Fieke Raaijmakers, Mijke Wassink, Nanne Ongena, and Pauline Fernandes. In addition, I would like to thank Nanne for his ideas and comments while I was working on chapter 7, and correcting the final version of the manuscript.

As part of this research project, I joined the Interuniversity Graduate School of Psychometrics and Sociometrics (IOPS). The courses that I took in this research school were very helpful and interesting. I would also like to thank all members of IOPS who gave any comments or advice at the IOPS conferences.

I would like to thank Miriam Fossen for her encouragements and kind advice. Finally, I thank my family and friends, whom I cannot name all in person, but you have certainly supported me.

1 Introduction

1.1 Survey interviews

The goal of survey research is to collect information that reflects actual behaviors, attitudes, and characteristics of people. This information is collected by means of measurements that are typically conducted with a sample of respondents in order to generalize to a population. Measurement can take place by means of self-administration (e.g., paper and pencil questionnaires or web surveys) or by means of an interviewer. According to Dijkstra and Van der Zouwen (1982), interviews have the advantage over self-administered questionnaires that some control can be exerted over the respondent's task performance, and survey concepts can be clarified by interviewers. Furthermore, they enable increased certainty that the intended respondent answers the questions, and yield more cooperation of respondents as compared to self-administered questionnaires.

A survey interview can be defined as:
a two-person conversation, initiated by the interviewer for the specific purpose of obtaining research-relevant information, and focused by him on content specified by research objectives of systematic description, prediction, or explanation (Cannell and Kahn, 1968, p. 527).

The systematic character of survey interviews is generated through standardized questionnaires. Furthermore, interviewers are instructed to conduct interviews in a standardized manner (e.g., read questions as written and probe non-directively, see Fowler and Mangione 1990). Although it is assumed that the concepts being measured have true values independent of the survey, in the process of collecting survey data, several types of measurement error can occur. Such errors might cause the respondent to answer questions inaccurately or imprecisely, or to answer a question with a different meaning than intended by the researcher. These errors may be related to the method of data collection, the measurement instrument, and interviewer and respondent characteristics (Sudman and Bradburn 1974). Several methods exist to obtain information about the quality of the responses, especially intended to improve questionnaire design, such as cognitive interviews, expert reviews, etc. (for categorizations and comparisons of several methods see Biemer 1988; Esposito et al. 1992; Krosnick 1999; Van der Zouwen 2001). In this thesis a method will be used that analyzes the verbal utterances of interviewer and respondent during the interview, in order to gain insight into the question-answer process and errors that may occur during this process.

1.2 Studying the interview process

Cannell, Fowler and Marquis (1968) state that within the interview itself, particularly in the behavior of the participants, we can find important causes of good and poor survey responses. In their description of the interview, Cannell and Kahn (1953) focus on behavioral patterns

and the interaction between the interviewer and respondent. As Cannell, Oksenberg and Converse (1977, p. 308) already noted, the more 'traditional' studies on interviewer effects are concerned with 'invariate' characteristics of the interviewers (e.g., their age or sex) and not with "the actual dynamics of the personal interaction between the interviewer and the respondent". The formal tasks of an interviewer are to ask questions, record answers, probe neutrally, and motivate respondents. Cannell et al. (1977) were interested in how interviewers differed from each other with respect to the performance of these tasks, and how this affected the motivation and performance of respondents.

The importance of studying the interviewing process has gained more and more recognition in the past 30 years. Although the first studies were primarily directed towards the behavior of the interviewer, in order to detect bad interviewer performance and its effect on data quality, it soon became apparent that the behavior of the respondent is equally important in understanding the question-answer process, and its effect on the quality of the eventual data.

The respondent's role is to give answers that are adequately formulated and accurate. Especially when respondents do not understand the task of answering standardized questions problems in the interaction may occur. Survey questions place high demands on respondents, and response errors may especially occur because respondents are not able or willing to meet these demands (Cannell et al. 1977).

The interviewer is an important factor in motivating respondents. Cannell and Kahn (1968) assume that scripted texts read by interviewers are not useful to motivate respondents. A script cannot anticipate the respondent's mood or need for explanation. It is more probable that spontaneous interaction enhances a higher motivation of respondents.

1.3 Interaction analysis

Interaction analysis entails a thorough examination of the verbal utterances during the process of question and answering. This analysis can give insight into the way information is exchanged between the interviewer and the respondent, and how the eventual answer is obtained.

The unit of analysis in interaction studies is usually the question answer sequence (Q-A sequence). Such a Q-A sequence consists of all utterances that belong to a single question. The Q-A sequence starts at the moment the interviewer asks a question and it ends by posing the next question, which indicates the interviewer has acknowledged the respondent's answer (Dijkstra 1993). These Q-A sequences can be analyzed with respect to the occurrence of behaviors that may negatively affect the outcomes of the interaction process. For example, when interviewers do not pose the question as worded, they may subtly change its meaning and thus influence the answer of respondents. Of course analyses can be expanded across Q-A sequences, taking the entire survey interview as a unit of analysis.

Both the interviewer and the respondent can cause deviations from the so-called 'paradigmatic' sequences. Schaeffer and Maynard (1996) introduced this term to indicate sequences that are perfect from a survey researcher's point of view. During a paradigmatic

sequence (or 'straightforward sequence', Sykes and Morton-Williams 1987) the interviewer poses the question as scripted and the respondent immediately gives an adequately formatted answer that is assumed to be appropriate.

The fact that many Q-A sequences deviate from the paradigmatic sequence indicates that answering a survey question is an interactive and co-operative process (Schaeffer and Maynard 1996). During this process both the respondent and the interviewer invest cognitive and social effort, and may influence the interaction in several ways. These influences can be problematic for the quality of the response obtained. We can observe interactions and try to identify problematic and non-problematic deviations from the paradigmatic sequence. Which deviations will be considered as problematic may depend on the researcher's point of view. For example, a researcher propagating strictly standardized interviewing, will not allow interviewers to use minor deviations from original wording, and may view elaborations of answers as problematic because they distract the respondent and interviewer from their task (e.g., Fowler and Mangione 1990). A researcher propagating standardization in a less strict way, may view minor deviations in question reading and elaborations of answers as harmless, or even as having a positive effect on the respondents' motivation and hence on the quality of the data.

Interaction analysis can be done *prior* to collecting data of interest, e.g., in order to pretest questionnaires. This may comprise an iterative process of pretesting and improving the questionnaire or interviewer instructions. Interaction analysis can also be done *after* actual data collection to identify measurement errors and to explain biases in the data obtained with interviews. Furthermore, a goal of interaction analysis after actual data collection may be to improve survey design in general, and add to the theoretical knowledge of question answer processes in the interview.

The latter approach is the goal of interaction analysis performed in this thesis. We aim to detect systematic problems that occur in the interaction between interviewers and respondents. We also aim to identify the sources of these problems, and evaluate whether such sources can be influenced in order to prevent problematic deviations.

1.4 Usefulness of interaction analysis

The results of interaction analysis can provide new theoretical insights and can also give suggestions for improving survey research measurement. The results of analysis of specific utterances, and relations with characteristics of questions, interviewers and respondents that cause problematic deviations can give clues for improving interviewer training and instructions for respondents during the interview. The frequency of occurrence of problematic deviations can be used to identify individual questions or series of questions that appear to be problematic for interviewers and/or respondents, and reasons as well as solutions for these problems can be suggested. Moreover, information about the occurrence of problematic deviations may indicate the quality of the data that is collected with a survey interview. The relation between the validity of responses and the occurrence of several problematic

deviations in interviews has been demonstrated in several studies (e.g., Belli and Lepkowski 1996; Dijkstra and Ongena forthcoming; Dykema, Lepkowski and Blixt 1997).

Interaction analysis can support or add to the results of other procedures to detect measurement errors. Sykes and Morton-Williams (1987) point out that it can only detect errors that are manifested interactionally. Although in a paradigmatic sequence by definition no manifested problems occur, other problems that are not directly observable may affect data quality. For example, respondents can answer in a socially desirable way, pursue satisficing strategies (Krosnick, 1991) or fail to let the interviewer know they had trouble in understanding a certain question. However, as long as non-paradigmatic Q-A sequences are available to a sufficient degree, interaction analytic studies provide a lot of information about problems that affect data quality.

So, the usefulness of methodological research, such as interaction analysis, depends on the extent to which non-paradigmatic Q-A sequences actually occur in regular surveys. As Cannell and Kahn (1968) note, we do not know to what extent studies on the reliability and validity of survey data are representative for general survey practice. A first bias they mention is that when no problems in measurements are found, this tends to be never reported. Lack of problems in methodological studies is considered a less interesting result. Furthermore, the data analyzed often comprises data from carefully designed surveys. As Cannell and Kahn state it “the careless or unsophisticated researcher is not likely to offer his data for methodological research and is still less likely to do such research himself” (Cannell and Kahn 1968, p. 540). From the first observation (only problems are reported) we might conclude that methodological research constitutes an over-estimation of problems in survey research. However, from the second observation we might conclude that the problems reported are only a tip of the iceberg. If so many problems in measurement can be reported about the best surveys, the regular survey is much more problematic than the surveys that are carefully scrutinized for measurement errors.

1.5 Overview of the thesis

At present, we do not know to a very detailed level how the process of questioning and answering actually takes place, i.e., when and why a Q-A sequence deviates from the paradigmatic one. In order to detect causes of problematic deviations, and how these deviations are related to other behaviors in the Q-A sequence, we need a systematic and efficient method to study interactions in survey interviews. Thus, in this thesis we aim to answer four research questions:

1. *What type of problems in interaction can be expected in survey-interviews from a theoretical point of view?*
2. *What is the most appropriate method to identify interactional problems in survey interviews?*
3. *Which problematic deviations from a paradigmatic question-answer sequence occur most frequently, which actors are mostly responsible for these deviations, and how are these related to other behavior in the question-answer sequence?*
4. *What theoretical explanations can be found for the occurrence of problematic deviations in question-answer sequences?*

Chapter two will address the first research question. In this chapter, the interaction between the interviewer and respondent is described from two theoretical viewpoints. The first is a conversational perspective. The second refers to the cognitive processes involved in the question answering process. Both perspectives provide clues for behaviors that are relevant to the study of Q-A sequences.

In chapter three, the second research question, concerning methods for the identification of problems in Q-A sequences, is addressed. Behavior coding as a method to detect problems in survey interviews is described. This is done by means of a comparison of different coding schemes, and the different coding procedures that can be applied. In chapter four, the coding scheme is described that is used in the empirical chapters in this thesis. The codes that are included are also compared to codes of other coding schemes.

The third research question is addressed in chapter five. By means of an analysis of existing interview data, it is established which problematic deviations occur most frequently and which actor most frequently causes the first problematic deviation in a Q-A sequence. Furthermore, the relation between problematic deviations and other behaviors in Q-A sequences is described.

In chapter six, hypotheses concerning a theoretical explanation for the occurrence of problematic deviations are formulated and tested by means of a non-experimental study, again using data from an existing survey, i.e., a survey conducted for quite different purposes. Finally, the hypotheses are tested in a field experiment, especially designed for this purpose. This experiment is described in chapter seven. In chapter eight the results of all preceding chapters are summarized, final conclusions are drawn, and recommendations are given.

2 Interaction in the survey interview from a cognitive and conversational perspective

2.1 Introduction

As we described in the previous chapter, in a survey interview the interviewer has multiple tasks, such as asking questions, recording answers, probing neutrally and motivating respondents to provide answers. The respondent has only one task, that is, answering the questions. How well both participants perform their tasks may depend on characteristics of the respondent, the interviewer, the questionnaire, and the social context. The tasks of the interviewer and respondent can be described from two different perspectives. The first one is a conversational viewpoint; a survey interview can be considered a two-person conversation. Thus, how communication takes place in ordinary conversations, may affect the way interactions proceed in survey interviews. This conversational perspective emerged from theory and research in sociolinguistics on survey interviews. The second viewpoint is a cognitive viewpoint; how respondents answer survey questions can be described from cognitive theories about information processing and memory. The cognitive perspective was developed from theory and research in social and cognitive psychology on survey interviews.

2.2 The standardized interview from a conversational perspective

Several studies have pointed out that the interview is an interactional setting in which interviewer and respondent communicate according to rules comparable to rules of ordinary conversations (Cicourel 1982; Clark and Schober 1992; Means et al. 1991; Schaeffer 1991; Schwarz 1996; Suchman and Jordan 1990). As Suchman and Jordan (1990) note, researchers interested in oral communication often take the ordinary conversation as a 'baseline' for their analyses, because the minimal requirements for orderly, mutually intelligible talk can be found within ordinary conversations. Reasons for problems in the flow of the interaction in a survey interview may therefore be found in comparisons of standardized interviews with ordinary conversations.

As Holbrook et al. (2000) point out, experimental studies on adapting procedures in survey research to conventions and norms governing ordinary conversations hardly exist. Schober and Conrad's (1997, 2000, 2002) studies are an exception to this rule. However, a lot of studies that examine interaction in the survey interview in a non-experimental way illustrate the incongruity between standardized interviews and ordinary conversations. Ahead of a review of such studies, we first need to define both types of communication.

2.2.1 Definition of a standardized interview

The standardized survey interview can be described as a conversation with the purpose of collecting valid and reliable data. The main goal of standardization is to collect data that is comparable across respondents by keeping the stimuli (i.e., the questions) provided to them constant. A first means of standardization is a questionnaire with a predetermined structure,

question order, and question wording. A second means of standardization are the instructions that interviewers must apply in interviewing. With standardized scripts for interviewer behavior the researcher attempts to create constant interviewer behavior.

Survey researchers assume that all Q-A sequences develop along the pattern of a paradigmatic sequence (see section 1.3, chapter 1). Several studies showed that when interviewers differ in the way they administer questions, they may affect data quality. For example, for changes in question wording, Schuman and Presser (1981) concluded that small changes in the wording of a question may be related to substantial differences in the distribution of responses. These results founded the basis of standardized interviewing (Beatty 1995). The basis of the techniques of standardized interviewing can be summarized by means of four techniques that Fowler and Mangione (1990, p. 35) give; reading the questions as written, probing non-directively after inadequate answers, recording answers without judgment, and being interpersonally non-judgmental regarding substance of answers.

2.2.2 Definition of conversation

Sacks, Schegloff and Jefferson (1974, p. 696) refer to "talking in interviews, meetings, debates, ceremonies and conversation" as 'speech exchange systems'. They suggest that "conversation should be considered the basic form of speech-exchange systems" (Sacks, Schegloff and Jefferson 1974, p. 730). The general character of conversations may also be derived from Slugoski and Hilton's (2001) definition of conversation:

> *We define "conversation" as an orderly, jointly managed sequence of utterances produced by at least two participants who may or may not share similar goals in the interaction. (p. 194)*

This definition addresses the coordinative character ('jointly managed') of conversations. As we will point out, this is an important characteristic that may set up a tension between ordinary conversations and standardized survey interviews.

Furthermore, the definition includes the aspect 'goals' that participants in a conversation may or may not share. Although Slugoski and Hilton do not elaborate on this aspect, we take it as an important aspect of conversations, especially usable to distinguish a conversation from a standardized interview. We will take the goal as a first point in an (far from complete) overview of differences between ordinary conversations and standardized interviews.

2.2.3 Goals and motivation of the participants

The difference between standardized interviews and conversations has largely to do with a difference in the goals and motivations of the participants. The goals of the participants, i.e., interviewers' and respondents' goals, are usually different from the researcher's goal, to collect valid and reliable data. From a negativistic point of view, both interviewers and respondents may be interested only in the compensation they receive for doing the interview (e.g., financial or non-monetary incentives), and therefore they may pursue the goal to finish

the interview as quickly as possible. However, interviewers and respondents may also mutually differ in their goals. Interviewers' goals may consist of doing their job well, getting satisfaction out of their job, and keeping their job. Respondents' goals may be to perform a citizen's duty, to find a way to express themselves, to perform an intellectually challenging cognitive task, or to have a chance of evaluating themselves (e.g., respondents who see the interview as a psychological test).

The goal of participants in conversations is often quite mutually alike. For example, unacquainted people waiting at a bus stop, accidentally talking, while waiting for a bus to arrive, may have the mutual goal to just pass time pleasantly. As appears from a study by Cannell, Fowler and Marquis (1968), specific goals of both the interviewer and the respondent may resemble the goals in ordinary conversations. For example, the main appeal interviewers reported about their job was "the chance to come into contact with other people". Furthermore, for the respondents, the second most given reason for cooperating in the survey was that "the respondent merely enjoyed being interviewed or enjoyed having a chance to talk to someone" (Cannell et al. 1968, p. 5).

Their reason for cooperating may take respondents to believe that the interview is similar to ordinary conversations, and thus may contribute to the interactional troubles that arise from this belief. In addition, they may be disappointed when a survey interview turns out to be different from the nice conversation that they expected to have. Suchman and Jordan (1990) address the issue of repetitious and depersonalized scripts that may discourage respondents to keep some "sense of involvement with, or responsibility for the interview responses" (Suchman and Jordan, p. 235). The repetitive character of survey questionnaires may also discourage interviewers, as Mathiowetz and Cannell (1980) conclude from their behavior coding study.

2.2.4 Rapport and involvement

In order to motivate respondents to provide information, the interviewer needs to establish a positive relationship with the respondent, which is generally referred to as 'rapport'. Rapport is often assumed to negatively influence standardization (Beatty 1995). In Cannell et al.'s (1968) study, it appeared that a positive attitude (mainly found for older respondents) toward the interview did not indicate that respondents had an accurate perception of their task. According to Hyman (1954) rapport may be a function of the degree of total involvement. Hyman distinguished two kinds of involvement of a respondent. The first is 'task involvement', which is the involvement with questions and answers, and may increase validity. The second is 'social involvement', which comprises the involvement with the interviewer as a person. Respondents can be biased by social involvement, for example because they tend to agree with interviewer opinions (i.e., they are seeking for interviewer's approval or avoid to offend the interviewer, see also section 2.2.8).

Weiss (1968) observed that respondents who were rated by interviewers as most "confiding" provided more biased information than respondents who were rated as "hostile". However, as the ratings concerned interviewer's evaluations of respondents, this refers to a

rather subjective measure. Unfortunately, Weiss does not elaborate on the specific instructions interviewers received for applying the ratings.[1] A respondent who is rated as 'hostile' by interviewers might in fact be a respondent who criticizes the survey instrument or the interview practice. Such a respondent therefore might approach the survey interview in a more serious way, and may actually be quite involved in the task (i.e., Hyman's 'task involvement'). In contrast, the nice and polite respondent who is just having a nice chat and therefore may be rated as 'confiding' may be less involved in the task but nevertheless involved in the conversation (i.e., Hyman's 'social involvement').

An interaction that illustrates how a critical respondent can demonstrate behavior that may be interpreted as hostile is included in Excerpt 2-1. The respondent in this interview was doing more than just trying to give adequate answers. He was also communicating his opinion about the questionnaire (see line 4). An interviewer might interpret such reactions as hostile, but it seems plausible that this behavior indicates task involvement and therefore may be positively related to adequate answers.

Excerpt 2-1 Critical respondents*

1.	I: And once again for weekdays: what is the total number of cups of water, coffee, tea and other non-alcoholic beverages that you usually drink on one day?
2.	R: On one day huh?
3.	I: Yes
4.	R:You should categorize this, this is too difficult for the average answers, I suppose, you cannot do that, the conversation will last way too long.
5.	R: Uh what could it be? What could it be?
6.	R:Ten
7.	I:Ten, Okay

*This Q-A sequence was slightly abbreviated from the original, taken from the Health Issues Survey that is described in chapter 7.

Interviewer behavior may also influence respondents' behavior. As Weiss (1970, p. 20) in a review suggests, the "important factor for securing valid answers is the respondents' understanding of his role as information-giver. Good professional performance by the interviewer, rather than personal comraderie, may do the job." A socio-emotional style of interviewing may, as Dijkstra, Van der Veen and Van der Zouwen (1985) conclude, be viewed as less efficient (but not necessarily less accurate) than a formal style. However, the effect of the immediately preceding interviewer behavior on the behavior of the respondent appeared to be much greater than the effect of the interviewer style.

[1] All that is reported about this rating is the five-point scale (including the options 'confiding', 'frank', 'equivocal', 'guarded' and 'hostile')

2.2.5 Participants

Ordinary conversations are not restricted with respect to the number of participants involved. Most conversational principles also do not depend on or differ with respect to the number of participants involved. However, in standardized survey interviews, interviewers are often instructed to see to it that the interview is done in a private setting, to prevent third parties (e.g., household members) from overhearing the interview. Although, as Tourangeau, Rips and Rasinski (2000) note, experimental studies failed to show convincing effects of the presence of local third parties on the quality of answers to questions about sensitive topics, such presence will certainly influence the flow of the interaction. As the Q-A sequence in Excerpt 2-2 shows, respondents may, *after* they have already provided an acceptable answer, request for clarification (line 6) based upon a comment of a third party (line 5).

Excerpt 2-2 Effects of the presence of a third party*

1. I: Uh which policy would you prefer with respect to Eastern Europeans who come to the Netherlands to live here?
2. R: Eastern Europe, pffh, well I would want to be stricter on that
3. R: Three
4. I: Three
5. P: You are discriminating
6. R: How should I see that, accept everybody?
7. I: Yes, eight, then it is good that they come here and three then it is like that uh you actually would want to stop them
8. R: Yes

*This Q-A sequence was taken from the European Social Survey Pilot data, that is described in chapter 6. The response options, as presented to the respondent on a show card, constituted an 11-point scale, ranging from '0, Stop everyone who wants to come here' to '10, Accept everyone who wants to come to live here'.

2.2.6 Topics

An important difference between ordinary conversations and the survey interview is the control of topics. According to Suchman and Jordan (1990, p. 233) the "central organizational feature of ordinary conversation is that who talks, and about what, is controlled from within the conversation by the participants". In ordinary conversations the topic will mainly determine whether partners will find it worth the effort to continue talking. This effect is strongest in accidental conversations (e.g., the earlier mentioned example of unacquainted people waiting at the bus stop).

The advance determination of topics may result in the danger of an uninterested interviewer or respondent. This was of greatest concern in the 1954 University of Michigan's Survey Research Center's '*Manual for Interviewers'*. Beatty illustrates this concern with a citation from the manual: "Each question should be asked in a manner implying that it presents an interesting topic, and that you are extremely interested in having the respondent's ideas on it" (cited by Beatty 1995, p. 151).

Furthermore, in ordinary conversations people often mutually exchange knowledge of a topic. One of the participants may have expertise on a particular topic, but often people share knowledge in the domain that is discussed. As Fowler and Mangione (1990) suggest, an interviewer is ideally not an expert on the topic of the interview. When interviewers have specialized knowledge on the topics of the interview, they may assume to know respondents' intended meanings of unclear answers. That knowledge may influence their probing behavior, and consequently responses obtained (Fowler 2002).

2.2.7 Cooperation

According to the principle of cooperation (Grice 1975), conversations take place by means of cooperation. This entails that speech partners communicate by adjusting their verbal behavior mutually. Grice elaborates his principle in four interpretation rules, conversational maxims. The maxim of *quality* drives speakers not to speak of things that are not true or insincere. The maxim of *quantity* prescribes that utterances are as informative as is necessary, but no more informative than is required. The maxim of *relation* prescribes utterances to be relevant for the conversation that they are part of. And finally, the maxim of *manner* prescribes utterances to be clear, and therefore ambiguous or unfamiliar terminology should be avoided.

An important reason for Grice to describe these maxims was to illustrate how people can, by deviating from the maxims, use *conversational implicatures*. A conversational implicature is an inference about the meaning of an utterance, which can be drawn because of the fact that the speaker is considered to be a cooperative communicator.

Schwarz (1996) gives a review of several examples that particularly illustrate the relation between the principle of cooperation and biases in understanding.[2] Because respondents belief that the researcher and/or the designer of the questionnaire is cooperative, they believe that all information offered by the researcher is relevant, that false presuppositions are not deliberately included in questions, and that general and specific questions can be judged as conversationally related. However, such effects only occur when respondents have reasons to assume that the communicator has knowledge about the issue and is willing to adhere to Gricean maxims (Schwarz 1996). In addition to its influence on understanding survey questions, the principle of cooperation will influence the interaction between the interviewer and the respondent in several ways. In several sections of this chapter (e.g., sections 2.2.12 to 2.2.15) we will give such examples.

2.2.8 Face strategies and Politeness

As Tourangeau et al. (2000) argue, politeness strategies are an important reason for misreporting in standardized surveys. In social interactions (and therefore in survey interviews as well), people will present a public image of themselves that Goffman (1967)

[2] Tourangeau et al. (2000) argue that it is not very likely that implicatures will always work out in unintended ways. As they state it, respondents "can't overinterpret *everything*" and: "nearly any effect of question wording can appear to be an implicature after the fact" (Tourangeau, Rips and Rasinski, 2000, p. 54).

referred to as 'face'. Both speaker and listener appear to treat their face with care. Someone can loose one's face, but also threaten the other's face.

According to Brown and Levinson (1987), face comprises a positive and a negative component. Positive face concerns the need to be appreciated and acknowledged by other people. This need can be fulfilled when people specifically address it with compliments or indications of solidarity. Negative face concerns the need for freedom of acting. This need can be fulfilled by avoidance (hence the 'negative' label) of addressing face threats (i.e., not posing a question at all, or posing it in a very polite way), which means that a speaker leaves more room for freedom of acting to the listener.

When people have to perform 'face threatening acts', they try to do this in tactful ways. A speaker may, in the most polite way of dealing with the face threat, not convey the message at all. A speaker may politely convey the message using an 'off record' strategy (e.g., using indirect ways of communications, such as conversational implicatures), or strategies that compensate for positive or negative face threatening acts (Brown and Levinson, 1987).

In line with these notions, a standardized interviewer in a standardized interview is threatening the negative face of respondents by asking them to cooperate in a survey interview, and by insisting on a choice among the response alternatives when respondents have given only general information. Furthermore, she may threaten the positive face of the respondent when asking sensitive questions. Questions may be reworded by the interviewer in such a way that the positive face threat is compensated (i.e., Houtkoop-Steenstra's examples of no-problem questions, see section 2.2.16). Negative face threat may also be compensated by question rewording (e.g., "Could you tell me what was your monthly income during the past 12 months" rather than "What was your monthly income during the past 12 months"). However, compensation strategies may have harmful consequences for the clarity of questions, and at least lengthen the question wording. As Bradburn and Sudman (1979) pointed out, when the question length increases, errors and variance in question reading are likely to increase. Cahalan et al. (1994) also found that long questions, especially the ones with qualifying statements yielded more problematic interviewer behaviors (such as variations in question reading), and problematic respondent behaviors (such as requests for clarification and qualified answers). However, Marquis and Cannell (1968) found that long questions also generated better answers.

2.2.9 Turn taking

Slugoski and Hilton's (2001) definition of conversation (see section 2.2.2) indicated that conversation entails a 'jointly managed sequence of utterances'. As Clark (1985) notes, one of the ways to enable such a jointly managed sequence of utterances is a system of turn taking that allows for smooth distribution of turns. Sacks, Schegloff and Jefferson (1974) describe the rules of turn-taking that are universal for conversations. Their article can be considered as a milestone for the history of Conversation Analysis (Mazeland, 2003). Sacks, Schegloff and Jefferson argue that a smooth interaction, avoiding simultaneous speech

('overlap') and silence ('gap'), is possible because of the reliance on the organization of turn-taking, which comprises a 'turn-constructional' component and a 'turn-allocation' component.

The turn-constructional component is used to communicate possible points of change in turn taking (i.e., '*transition relevance places*'). A speaker can use various unit-types to construct a turn, such as anything in between constructions of a complete sentence or only one word. Turns are constructed out of one or more 'turn constructional units' (TCU's). A listener can put up expectations upon the completion of the ongoing TCU by means of syntactic, pragmatic or prosodic completeness (Mazeland, 2003).

The 'turn-allocation' component is used to select the next speaker. In a rather explicit way, this selection of speakers can be done by the production of the first part of an 'adjacency pair'. Such a pair consists of two utterances that are in a certain way related to each other, and are placed one after another, by different speakers. Since a question can be considered as the first part of such a pair, this is in survey interviews the most important turn-allocation procedure. By producing a question, the current speaker selects the next speaker (i.e., the person who is supposed to answer the question). Thus, the answer is the second part of the 'adjacency pair' and must be produced by the person whom is addressed in the first pair-part (Schegloff and Sacks 1973).

Turn-taking in standardized interviews

The turn-taking system in standardized interviews may be viewed as a simple sequence of adjacency pairs, i.e., just questions and answers. However, as Houtkoop-Steenstra (2002) illustrates, scripted interviewer's utterances comprise more than questions alone. Examples of various components mentioned by Houtkoop-Steenstra are summarized in Table 2-1.

Table 2-1 Components of Interviewer Scripts

Component:	Abrev.:	Indication of:	Example:
Action Projection	APC	Type of action	'I will now ask some questions'
Question Target	QTC	Topic of question	'…these will be about X'
Question Specification	QSC	Definition	'…by X we mean…'
Question Delivery	QDC	Question	'How often do you do X'

These multiple utterances need to be communicated *within* one turn (i.e., a multi-unit turn). When the transition relevance place of such a multi-unit turn is not projected clearly, problems in the interaction may result.

As might be expected, respondents will provide an answer as soon as they have heard the question delivery component (QDC). However, many survey questions are structured in such a way that definitions and specifications are to be read *after* the QDC. When respondents start answering as soon as this question component is delivered, the specifications are likely never to be heard by respondents, and the consequence may be that "the original survey question is

interactively transformed into a question with a different meaning" (Houtkoop-Steenstra, 2002, p. 249).

Another multi-unit turn that is often used as a question structure in surveys, is that the response alternatives are read *after* the question proper. Respondents are likely to provide an answer before the interviewer has finished reading the alternatives. As a consequence, respondents are not fully informed about the response alternatives, and are more likely to provide answers that are not formatted according to the alternatives as scripted (see also section 2.2.14).

It is difficult to prove that interruptions generate inadequate responses. As long as the respondents do not say anything that contradicts earlier answers, interruptions may not threaten the quality of the responses. However, when respondents are not informed of all specifications or response alternatives, answers generated in this way cannot always be trusted. As Schaeffer (2002) points out, to reduce the chance of interruptions, a long question must be transformed into several shorter questions, or the question delivery component must be placed at the end of an item.

From several behavior coding studies (Bates and Good 1996; Blixt and Dykema 1995; Burgess and Patton 1993; Prüfer and Rexroth 1985; Snijkers 2002; Sykes and Collins 1992) it appears that questions are frequently interrupted by respondents. Interrupting question reading is also shown by Lepkowski, Siu and Fisher (2000, p. 3): "to be a function of exposure to questions that exhibit wording that is lengthy or contains numerous clauses that qualify the topic of the question".

Next to the construction of survey questions, construction of probes may benefit from the knowledge about turn taking. As Stax (2004) and Houtkoop-Steenstra (2000) both note, in survey research, specific directions of how a probe should actually be worded are generally never given. According to standardized survey practice, it is only indicated that the probe should contain all (or at least all relevant) response alternatives, and it should be worded in a neutral way, not suggesting a particular response alternative.

Stax (2004) suggests that an 'x or yz' format (such as "*Do you fully [x] or partly [y] agree [z]* ") fulfills the requirement of a probe format that is not likely to be interrupted, because it signals that more response options are underway. In this particular format, the response options are formulated in a construction that first lists the 'modifiers' ('fully' and 'partly') and ends with the predicate ('agree'). The frequently used 'xz or yz' format (i.e., "*Do you fully [x] agree [z] or partly [y] agree [z]?*") consists of 'fully elaborate units'. The fact that the response options are presented as fully elaborate units may signal a transition relevance place (see second paragraph of this section) too early. Respondents may infer that only this one response option is being presented. Consequently, this format is much more sensitive to interruptions by respondents, as Stax also showed empirically.[3]

[3] Although conversation analytic studies hardly ever mention frequencies, it is interesting that Stax (2004) makes some effort to present quantifications. From her study it appeared that a '*little under half* 'of the 'xz or yz' formatted probes are interrupted before a second option is uttered by the interviewer. In contrast, '*most of*' the uninterrupted probes appear to be formatted according to the 'x or yz' format.

2.2.10 Third turn options

An adjacency pair may be supplemented with a 'third turn' by the speaker who also produced the first part of this pair (Heritage 1984). With this third turn, the first speaker may display reception of the second pair-part. As Houtkoop-Steenstra (2000) illustrates, in this third turn, receipts may be produced that indicate surprise, interest (i.e., assessments) or satisfactory termination of the sequence (i.e., acknowledgments). However, in standardized interviews, interviewers are only allowed to give neutral receipts, because assessments are assumed to influence respondents. Nevertheless, interviewers may use assessments to display a personal orientation (Houtkoop-Steenstra, 2000).

Receipts that are designed by means of full or partial repeats of the second pair-part, are common (and preferred) in survey interviews, but rarely occur in ordinary conversations. Houtkoop-Steenstra (2000) explains this uncommonness in ordinary conversations, with the fact that such neutral repetitions only display a proper hearing, but do not necessarily indicate a proper understanding of the second pair-part. In standardized interviews all that the interviewer needs to understand about the second pair-part is the adequacy of the respondent's answer, which is dictated by the response alternatives in the questionnaire. The main functions of this repetition are to indicate perception, and to check whether the correct response is recorded. However, perception can also be performed by means of a minimal response like 'uhuh' or 'yeah'. Therefore, when the interviewer repeats the respondent's answer, "the interviewer shows the respondent that [s]he is temporarily engaged in a non-conversational activity" (Houtkoop-Steenstra, 2000, p. 26). Furthermore, as interviewers, after an adequate answer, have to perform some action to record the answer (write it down or find and press a key) a repeat is an effective way to fill the silence while they are recording the answer.

2.2.11 Preference for agreement

The first part of an adjacency pair often implies the content of a second part. This content is governed by a preference for agreement. Disagreeing utterances are often preceded by hesitations, qualifications, or even by an initial agreeing response that is subsequently changed into the disagreeing response (Mazeland, 2003).

Although interviewer's suggestions are not necessarily correct, the preference for agreement may cause the respondent to agree with suggestions (Houtkoop-Steenstra 1994). Respondents who do not agree with the suggestion can start their non-preferred utterance with a hesitating 'yes'. Interviewers who, after this 'yes', immediately proceed with the next question disallow the respondent to change or qualify their answer. Smit, Dijkstra and Van der Zouwen (1997) confirmed in an experimental study that respondents indeed often accept interviewer's suggestions. Therefore, suggestive probing may be considered as a serious problem.

However, we may wonder whether suggestive probing does occur often, and why interviewers probe suggestively. As Van der Zouwen, Dijkstra and Smit (1991) point out,

suggestive interviewer behavior hardly ever occurs at the beginning of the Q-A sequence, that is, when the interviewer reads the question from the questionnaire. It appears that suggestive probing especially occurs after some other fault in the interaction. For closed questions these faults often comprise interruptions: the respondent answers the question before the interviewer has read the complete list of answer alternatives.

2.2.12 Audience and recipient design

In ordinary conversations, utterances are adapted to specific recipients (speaking to a child involves different language use than speaking to an adult) and to specific situations ('recipient design', Houtkoop-Steenstra 2000, Suchman and Jordan 1990). In survey-interviews, however, question wording is determined in advance, and usually designed for a large and heterogeneous group of recipients, adapted to all possible circumstances ('audience design', Houtkoop-Steenstra 2000). As Suchman and Jordan (1990) argue, this often results in awkwardly structured questions that are difficult to read. Furthermore, it is impossible to account for all possible circumstances. Therefore, interviewers have the tendency to breach standardization rules, and read questions in their own adapted wording. Dykema et al. (1997) found that 'tailoring questions to a specific respondent's situation', omitting parts of a question that perhaps are inapplicable, or omitting parts that respondents already understand, appeared to increase the accuracy of the answers obtained. Apparently, tailoring questions did not change the question's meaning in an unintended way, which is the main reason for not allowing changes in question wording (Fowler and Mangione 1990).

The adaptation of utterances in ordinary conversations concerns also 'common ground' that is built during the conversation. Common ground is a source of information that can be used to adjust verbal behavior during a conversation according to the principle of cooperation; participants do not ask about things that are already in the common ground. Adherence to exact question wordings may result in awkward situations when an interviewer needs to ask for information that the respondent already spontaneously provided. As Houtkoop-Steenstra (2000) illustrates, an interviewer can solve this interactional problem with 'self-repair'. Interviewers often accompany a redundant question with a provisional answer (i.e., they in fact produce both the first pair-part *and* the second pair-part of an adjacency pair), or interviewers give remarks like "You've already said it but I have to ask". Such interviewer's 'self-repairs' indicate that the question is "retrospectively redefined by the interviewer as a case of reading a scripted line" (Houtkoop-Steenstra, 2000, p. 77). Thus, the interviewer signals the respondent that she is reading 'audience-designed' questions (or 'depersonalized' questions, Suchman and Jordan 1990).

Interviewers may also try to solve this interactional problem by not posing the redundant questions at all. However, the interviewer may overlook specific terms of questions or specific situations that the respondent did not report. Therefore, it is important that the information is verified rather than simply skipping questions and filling in something.

To conclude, allowing interviewers to change question wordings into recipient designed questions might solve interactional problems, and it may also contribute to the establishment

of a personal relation between the interviewer and respondent (rapport). Recipient designed questions may falsely signal respondents about what is expected of them. For example, interviewers may add phrases like 'on average' or 'as an estimation' to behavioral frequency questions. As a result, respondents may assume that general (i.e., imprecise) answers are sufficient (a problem that will be discussed in section 2.2.14). Since the interviewer, in adapting her questions to the respondent, has shown that she is paying attention to all respondent's statements, the respondents may assume that the interviewer will derive the exact answers from the imprecise answers provided by them.

2.2.13 Repair and understanding

In ordinary conversations, participants will try to solve problems, by initiating 'repair' before they continue the ongoing conversation. This means that participants can clarify ambiguous constructs. The interactive process of 'grounding' entails that in ordinary conversation references are only understood when the speaker and hearer both agree that understanding has been achieved (Schober 1999).

In ordinary conversations, indications that the hearer did not understand an utterance of the speaker *may* be ignored by the speaker, refusing to repeat his/her utterance (Churchill 1978, p. 108). In standardized interviews, ignoring misunderstanding is standard practice, as interviewers are prohibited to clarify the meaning of questions, in order to avoid interviewer variability. Interviewers may systematically differ in the way they clarify survey questions, and therefore it is assumed to be better to let the interpretation of survey questions entirely up to the respondent. The problematic aspect of interviewers providing clarifications after requests from respondents is that, as Moore and Maynard (2002) point out, the interviewer and respondent 'collaboratively modify question wording', whereas the other respondents will not receive this modified question.

From analyses of interview interactions, Moore and Maynard (2002) concluded that, in case of clarification proffers (i.e., utterances in which the respondent offers some question interpretation, like 'Does margarine count as butter?'), interviewers were twice as likely to respond in an unstandardized way than in case of explicit requests for clarification (e.g., 'What do you mean with butter?'). Clarification proffers not only indicate the source of the problem in understanding (e.g., the definition of butter) but also comprise an offer of candidate clarification (e.g., inclusion of margarine in the definition of butter). When respondents use these clarification proffers, it is very easy for interviewers to participate in modifying survey questions, because all that may be needed to reply is a short acknowledgement (Moore and Maynard 2002).

Explicit requests for clarification occur relatively rare in survey interviews (Schaeffer and Maynard 2002). Actions that mark but not specifically address respondents' problems are more common. For example respondents' reports (which are provisions of potential relevant information) and hesitations can also be used to identify problematic questions (Schaeffer and Maynard 2002). For example, in response to the question "Do you think of yourself as a Republican, Democrat, Independent or something else", a respondent may say "I suppose I

vote Democratic most often". By giving such reports, respondents try to leave the interpretative judgment involved in answering the question to the interviewer (Moore 2004).

As Moore argues, with these reports, respondents avoid to explicitly negotiate the problem, making them more efficient (from the viewpoint of the respondent) than explicit requests for clarification. The result may be that standardized interviewers, having learned not to substantially clarify survey questions, may be more likely to infer an answer, rather than to collaboratively negotiate the judgment.

The reason for occurrence of these reports may be that respondents are actually discouraged to explicitly request for clarification, as they know the response will be a standardized phrases like 'Whatever it means to you' (a 'WIMTY' response, Moore, 2004). Furthermore, a report is also a more efficient face-keeping strategy: respondents do not explicitly request clarification (avoiding a negative face-threatening action), nor do they admit they have trouble in understanding (avoiding a positive face-threatening strategy), whereas a request for clarification does both.

Paradoxically, the respondent's strategy not to explicitly address the problems in understanding, poses a face threat to the interviewer. According to standardized survey practice, an interviewer should ignore reports, and insist on a choice among the response alternatives. However, interviewers will probably not like to be viewed as ignorant interviewers who do not understand the respondent's troubles in answering a survey question (avoid threatening their own positive face). Moreover, they do not want to bother the respondent with neutral probing (avoid threatening the respondent's negative face).

In Figure 2-1, a model of respondent and interviewer behavior related to question comprehension is depicted. As we noted in section 2.2.7, the principle of cooperation (a) will influence question comprehension (b). Depending on the respondent's conversational strategy (c) a respondent may perform actions that explicitly (d), implicitly (e) or give no signal (f) of problems in understanding. For example, when respondents have high task involvement, they are more likely to explicitly request for clarification (g), request repetition of the question (h) or give 'clarification proffers' (i). Respondents' experiences with WIMTY's may cause respondents to more implicitly address the problem with reports (j). Other actions that mark a respondent's problem and therefore enable (but not necessarily require) interviewers' reactions are hesitations (k), imprecise answers (l), and comments (m) Respondents can also give precise answers, which give no indication of problems at all (n).

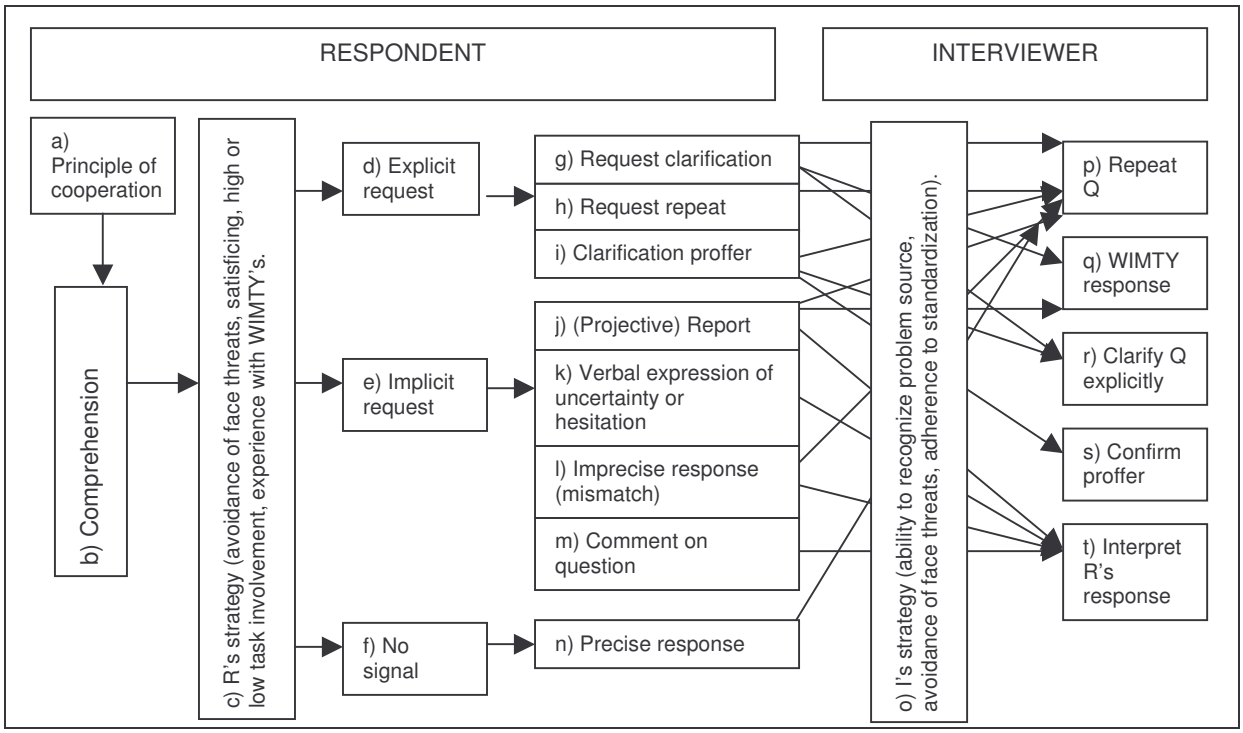

Figure 2-1 Model of respondent and interviewer behavior related to question comprehension

In a reply to the respondent's utterances that mark uncertainty of meaning, interviewers have several options to more or less explicitly clarify question meaning. This reply not only depends on the explicitness captured in the respondents' utterances (g-n), but also on the conversational strategy of the interviewer (o). According to standardization rules, interviewers are only allowed to repeat questions (p) or give a 'WIMTY' response (q). However, interviewers can of course also clarify the question (r), and in case the respondent offered some potential question interpretation in a clarification proffer, the interviewer may confirm this proffer (s). Implicit requests of respondents are usually provisional answers; in principle it is possible to infer the meaning of those answers (t). Whether they choose to do this may depend on their ability to recognize the problem source and their commitment to adhere to standardization rules (which is of course related to the extent to which they have been trained adequately). The avoidance of face threats may also play a role; face threats may easily be avoided with direct interpretation of the respondent's answer. After all, respondents have done their best to arrive at an answer, so why bother them with probing or clarifications. So here we have a problem: interviewers are not allowed to clarify question meaning, and face strategies may trigger interviewers to interpret the respondent's answer, and thus create inconsistent measurement, because interviewers are not likely to interpret respondent's answers in the same way. The reason for not allowing interviewers to clarify question meaning originates from a standpoint of offering standardized stimuli to respondents, but does not result in standardization of meaning.

These considerations of clarification of ambiguous concepts were the basis of Conrad and Schober's studies (Conrad and Schober 2000; Schober and Conrad 1997; Schober and Conrad 2002). They experimentally compared standardized interviewing techniques with conversational interviews (or as they labeled it, the 'collaborative approach'). It appeared that response accuracy is higher when interviewers have the possibility to provide clarification, than when they use strictly standardized interviewing techniques.

However, this result was at a considerable price; conversational interviews took much longer than standardized interviews (varying from 80% to 300% of standardized interviews). The success of the collaborative approach also depends on the extent to which interviewers are able to recognize implicit requests for clarifications as such. As Conrad and Schober (2000) indicate, 96% of the clarifications were given when it was not explicitly asked for. In some cases the clarification was prompted by a request to repeat the question. However, in most cases clarifications were given even though the respondent did not seem to have given explicit evidence that they needed clarification.

Although Schober and Conrad (2002) claim that their interviewers used the 'full resources of ordinary conversation' (as is also recommended by Suchman and Jordan, 1990), this of course has to be seen within the specific framework of the standardized interview. Unless we analyze all relevant specific verbal behaviors (word use, grammatical structures, sequential organization etc.) we do not know which characteristics of conversations are introduced when interviewers are allowed to deviate from their script. The extent to which the full resources of ordinary conversations can be used in survey interviews is limited by the fact that interviewers have a formal task and are instructed to collect specific information.

Repair issues have so far been limited to factual questions. Moore argues that in case of factual questions, respondents are more likely to use reports to put off judgment, than in case of "subjective" questions. He states that, in the latter case, "respondents themselves are always the ultimate authority" (Moore 2004, p. 60). However, as reports occur because the respondents have some kind of interpretation problem (Schaeffer and Maynard, 2002, p. 272), attitudinal questions may cause interpretation problems just as well as factual questions. The Q-A sequence in Excerpt 2-3 illustrates that respondents do not view themselves as the ultimate authority. The respondent (in line 2) not only gives a report, but also literary asks the interviewer what to answer. From the report in line 2 the interviewer may infer that the respondent agrees with the assertion, but from the report in line 5 this opinion is moderated. Eventually, after some discussion (not included in the excerpt) the respondent chooses, based upon an interviewer's suggestion, the response category 'agree'.

Excerpt 2-3 A report for an attitude question*

1.	I: The government should take care that there are smoke-free cafes and restaurants for the people.
2.	R:I am an anti-smoker so what should I say?
3.	I: Aha
4.	I: Do you consider yourself neutral or do you agree or strongly agree?
5.	R:I always say if someone wants to smoke I won't stop them
6.	I: No
7.	R:But I don't like to have it around me

* This Q-A sequence was slightly abbreviated from the original, taken from the Health Issues survey that is described in chapter 7. The five response options as presented in an introductory statement preceding the battery of three assertions were: strongly agree, agree, neutral, disagree or strongly disagree.

2.2.14 Restrictions in preciseness of answering

In an ordinary conversation it is often not necessary to answer questions with detailed accuracy. The Gricean maxim of quantity even prescribes speakers to be no more informative than necessary. In survey interviews, closed-ended questions with non-negotiable alternatives make up the majority of the questions posed. When respondents in survey interviews are asked how many days a week they watch television, they may think they are being cooperative when they answer "Most days" instead of exactly defining the number of days. However, such an answer is not directly codable by the interviewer because it does not match one of the fixed alternatives (i.e., a *mismatch* answer).

Prüfer and Rexroth (1985) provide three reasons for problems in response formatting that occurred in the interactions they studied: a lengthy battery of items, items with difficult content, and a non-visual verbal scale of four response categories. As a solution for the last problem, a visual presentation of the response categories (on a show card) might stimulate respondents to answer precisely. As Smit (1995) ascertained, especially in case of semi-open questions (also referred to as 'field coded' questions, see Houtkoop-Steenstra 2000) respondents have difficulties in adequately formatting the answer.

When the interviewer has to fill in a score (e.g., exactly defining a number of days in case of the question 'how many days a week do you watch television'), she has to probe until the respondent replies with such a specific answer. Because of the principle of cooperation, the interviewer is likely to probe in a suggestive way. A strictly non-directive probe, i.e., offering all response alternatives, may signal that the interviewer is uncooperative. By offering one or only a few alternatives that are warranted by the respondent's first answer the interviewer will not only signal that she was indeed paying attention to the respondent's utterance, she also makes the respondent's job a little bit easier (i.e., avoiding a negative face threat). However, the interviewer may not always be able to accurately determine the relevant range of answers, as may be derived from Excerpt 2-4. From the respondent's answer in line 2 the interviewer infers that 'once a month' might be an appropriate answer. However, the respondent corrects this suggestion (which might have been less likely the case when no show card was used). In line 4 the respondent gives another mismatch answer. Although in

line 6 the interviewer suggests another response alternative, the respondent in line 6 yet again gives a mismatch answer.

Excerpt 2-4 Problems in formatting the response*

1.	I:	Yes how often do you use uh do you use the Internet, E-mail or uh the World Wide Web?
2.	R:	Uhmmm, well I just said I just started, so that is not too often, no
3.	I:	Not often, but once a month?
4.	R:	No no no, that is more often
5.	I:	Multiple times a week
6.	R:	That must be twice a week
7.	I:	Couple of times a week

*This excerpt concerns a Q-A sequence taken from the European Social Survey as described in chapter 6. The response options, as presented to the respondent on a show card, were: 1 Every day, 2 Multiple times a week, 3 Once a week, Multiple times a month, 5 Once a month, 6 Less often, 7 Never, 0 (Don't know).

Interviewers may also avoid probing and use their own interpretation of the respondent's mismatch answer to arrive at the appropriate alternative. For example, if the interviewer in Excerpt 2-4 would have but just scored '1 month' right after the respondents' first answer in line 2, without any probing. Dijkstra and Van der Zouwen (1988) labeled this kind of behavior with the term 'choosing'. In that way, the interviewer rather than the respondent decides what response alternative is appropriate.

Although questions may generate no overt problems in a survey, other methods, such as cognitive interviews, nevertheless may reveal problems with respect to the requirement that exact and precise answers should be given. Beatty (2004) argues that standardized interviews suppress the expression of explicit problems. From this suppressing character of standardized interviews we may conclude that respondents select response alternatives and do not request for clarification because they are encouraged to restrict themselves to providing answers, and are discouraged to elaborate when they are unable to answer. Therefore only respondents who are truly involved in their task of giving informative answers, but also assertive enough to complain (and less likely to avoid negative face threats), will alert interviewers of the problematic character of questions.

It is also possible that in cognitive interviews, this very situation (i.e., its less standardized character) and not the specific difficulty of the questions, causes respondents to give imprecise (i.e., mismatch) answers. As Beatty states it: "the possibility that the conversational tone of the interaction discouraged participants from providing codable responses could not be dismissed out of hand" (Beatty 2004, p. 49). However, based upon observations of actual telephone survey interviews, Beatty concludes that imprecise answers are not unique to the cognitive interview situation. From his observations it seemed most plausible that respondents indeed produced a high amount of imprecise answers in the telephone interviews, but this was not observed from the eventual scores because interviewers used probing or their own interpretations effectively in order to obtain codable answers.

2.2.15 Elaboration

Even when respondents are able to give a codable response, they may cause interactional problems when they wish to explain or justify their response. Participants in ordinary conversations are free to digress and tell stories. Houtkoop-Steenstra (2000) notes that, therefore, respondents in survey interviews may provide information that is expected from a conversational point of view. Especially in case of yes-no questions, the answer is hardly ever formulated as a response that contains only a 'yes' or a 'no'. Molenaar and Smit (1996), for example found that in 47% of the cases, answers to yes-no questions were extended, with voluntary explanations for their answers. Such elaborations may cause interactional problems, because the additional information not only lengthens the interaction, it also distracts the respondent and the interviewer from their task.

2.2.16 Word order and the principle of optimization

In ordinary conversations participants tend to formulate questions optimistically, inviting optimistic, normal or 'no-problem' responses, a strategy Heritage (2002) labels as 'the principle of optimization'. Houtkoop-Steenstra (2000) gives some examples of questions in ordinary conversations in 'normal' wording (e.g., "Did you sleep well?" or "Is everything all right?") and in awkward wording (e.g., "Did you not sleep well?" or "Is anything wrong?").

From Houtkoop-Steenstra's (2000) analyses of ten interviews it appeared that interviewers often change a question with multiple response categories into an optimistic yes-no question (i.e., by suggesting the most optimistic response option). In this way, the interviewer creates a less complex question than the original question. Interviewers systematically picked the most positive response option to reformulate the question.

The optimistic order of alternatives is in correspondence with the convention that was addressed in a study by Holbrook et al. (2000). In this study the negative effects are illustrated of violating conversational conventions with respect to response order in survey interviews. Holbrook et al. argue that response alternatives that are ordered as different from word order according to conversational conventions may surprise respondents. For example, respondents expect an affirmative response alternative to be offered before the negative one, in case of dichotomous response alternatives (i.e., 'Do you agree or disagree' rather than 'Do you disagree or agree').

Therefore, they state that "at the very least, such a violation is likely to be momentarily distracting, pulling some cognitive attention away from simply answering the question, because one registers (even if unconsciously) the unexpected violation of conventional ordering." (Holbrook et al. 2000, p. 469). Holbrook et al. conclude that unconventional response order should be avoided, because it complicates the respondents' task, as it yields slower, less predictable, and presumably more erroneous responses, and an increased number of irrelevant thoughts. The study did not involve interviewers, but it seems reasonable to assume that interviewers as well may be distracted by unconventional word orders and as a consequence will produce more errors in question reading.

2.2.17 Word use

Ordinary conversations might be distinguished from standardized surveys by the specific words that are used. Researchers in the social sciences have their own expertise, and therefore differ with respect to their level of abstraction as compared to their general research population. Houtkoop-Steenstra (2000) for example, discusses an example of the question "What is currently your main activity". As she argues, 'main activity' is a typical example of a 'research-theoretical concept'. Although researchers also often include 'being unemployed' in this category, respondents are unlikely to consider being unemployed as a main activity.

As Houtkoop-Steenstra illustrates, when a question aims for categorical answers, it is very likely that respondents reply too specifically (i.e., mentioning specific instances within categories). Questions can be worded best in terms that respondents are likely to think of. For example, respondents are more likely to think of names of museums (e.g., 'Rijksmuseum', 'Stedelijk museum') than of categories of museums ('17th century art museum', 'modern art museum', etc.). Pretesting questionnaires, with methods such as cognitive interviews or focus groups (e.g., see Schaeffer and Dykema, 2004) may be useful to establish appropriate respondent terms.

Furthermore, whether or not respondents will actively use the formal words of a survey may be a matter of involvement. According to the coordination-engagement hypothesis (Niederhoffer and Pennebaker, 2002), the more the participants are actively engaged in a conversation the more verbal and nonverbal coordination is expected. Verbal and nonverbal coordination can be understood as imitation of the manner of speaking, and use of specific words, gestures, accent, etcetera of the other party. Engagement may occur in a positive way (e.g., participants are agreeing with each other) but also in a negative way (e.g., participants who are actively quarreling). Lack of engagement means that participants are simply not engaged in the conversation, i.e., they are not paying full attention (Niederhoffer and Pennebaker, 2002).

Therefore, a measure of linguistic style matching, which takes several linguistic variables (such as use of the same words, word size, grammatical tense, negations etc.) into account, can be used as a measure of conversational engagement. The level of 'conversational agreement' might be relevant for survey interviews as well. Respondents who are actively engaged in the interview will, as a result of their higher verbal coordination, be more likely to copy the interviewers wording. This means that they will have fewer difficulties in formulating answers exactly according to scripted response alternatives.

2.2.18 Summary

Standardized interviewing originates from a quantitative research tradition that is focused on standardized interviewer behavior to obtain reliable measurement. In contrast, non-standardized interviewing originates from qualitative researchers (sociolinguists and anthropologists), and is focused on interpretative measurement and validity. However, both have the goal to improve measurement of data collection by means of survey interviews.

In the discussion of differences between standardized survey interviews and ordinary conversations we have discussed the difference in topic control, the participants, and their goals. A respondent may believe that an interview is like an ordinary conversation. In a survey interview, respondents also use a system of turn taking similar to ordinary conversations. When a question or probe does not adequately project a transition relevance place (i.e., a possible point of change in turn taking), respondents may interrupt relevant the question or probe. Therefore question designers have to make sure that questions are not too long, and the question delivery component is presented last. Interviewers must make sure that probes do not contain fully elaborate response alternatives (i.e., asking 'do you fully or partly agree' rather than 'do you fully agree or partly agree'). Furthermore, questions in survey interviews are designed to account for all kinds of different situations, which may make them complex and difficult to read aloud. Respondents may also spontaneously provide information about questions to be asked later, which makes asking the question awkward. Face strategies and politeness may also influence interviewer's commitment to verbatim question reading and probing behavior (for example, interviewers may try to be polite with additions to questions such as 'May I ask you…?').

A respondent has several options to deal with ambiguous question meaning. These options differ with respect to the extent to which they explicitly address the problem. For example, respondents may explicitly request for clarification, or give a report, thus implicitly expressing a need for clarification. This difference will have consequences for the extent to which the interviewer will be able to recognize the problem source and to deal with them in a standardized way. Finally, conventions in word use and word order may be used to facilitate cognitive processing and to deal interactionally with questions. Uncommon word orders, which do not start with the most optimistic word first (i.e., 'do you disagree or agree?'), appear to disrupt cognitive processing of survey questions (Holbrook et al. 2000). This relates to the topic of the next section, in which we will discuss how cognitive processes influence the verbal interaction in the interview.

2.3 The standardized survey interview from a cognitive perspective

Answering a survey question involves several cognitive steps, as described in a well-known model of survey response (Tourangeau et al. 2000). The four steps described in this model are: (1) interpreting the question, (2) retrieving relevant information from memory, (3) forming a judgment from the retrieved information and (4) formatting the response.

Although it can be argued that cognitive processing does not take place in a nicely coordinated sequence of subsequent steps (see section 2.3.4), it is more convenient to think conceptually of separate tasks, and therefore we will deal with most of these tasks in separate sections, and explain how these cognitive tasks influence the interaction between the interviewer and respondent. It is difficult to separate retrieval and judgment with respect to their influence on the interaction. Therefore, these two tasks will be discussed within one section. After the discussion of the four cognitive tasks we will provide additions to Tourangeau et al.'s model of survey response.

2.3.1 Interpretation of a question

As Sudman, Bradburn and Schwarz (1996) suggest, within the interpretation of a question, a distinction can be made between understanding the contents and understanding the meaning of a question. A problem with respect to understanding the content of questions will occur as a consequence of lexical or structural ambiguities, for example because a foreign word, or an unconventional word or word order is used in the question. Not understanding the contents of a question is likely to result in a problem in understanding the meaning of a question, but not necessarily. It is also possible that a problematic part of the contents of a question only causes disruptions in processing the question wording, as was shown in Holbrook et al.'s (2000) study (see section 2.2.16). Problems with respect to understanding the intended meaning of questions are also relevant in ordinary conversations, and therefore we already discussed the consequences of problems in understanding questions for the interaction in section 2.2.7, and section 2.2.13 . In Figure 2-1 we gave a summary of the behavioral options of respondents and interviewers to deal with problems in question comprehension.

2.3.2 Retrieval and judgment of relevant information

When respondents retrieve information concerning autobiographical facts, they are not likely to follow a strategy to recall and count all relevant behaviors in a reference period (Schaeffer and Presser 2003; Schwarz and Oyserman 2001; Tourangeau et al. 2000). As Schwarz and Oyserman (2001), in their review of the most common procedures to improve retrieval of relevant information from memory note, respondents are unlikely to have detailed accounts of individual events stored in their memory. This is especially the case for frequent behaviors, and the quality of the information stored will decline the more distant in time events occurred. Even for important or unique events, memory decreases over time.

Tourangeau et al. (2000, p. 146) give a clear overview of strategies that respondents may adopt when answering frequency questions. Two of those strategies are not likely to be visible in the interaction; i.e., when respondents have exact tallies available or when they provide direct estimations based upon general impressions. When respondents use one of those strategies, they are likely to immediately produce an 'acceptable' answer.

The kind of strategy that survey researchers often hope for (or even assume), is that respondents recall each relevant event, and enumerate all events to get their answer ('recall-and-count' or 'episodic enumeration'). It will be clear that this enumeration will take some time. Respondents may start to verbally express their enumeration, to show the interviewer that they are busy with processing and to prevent that interviewers start repeating or clarifying the question because of long silences. Because these verbal expressions may trigger the interviewer to respond to these utterances , they may be problematic for the quality of the response obtained, as is illustrated in line 3 of Excerpt 2-5. In this line, the interviewer interrupts the respondent's enumeration, and suggests an answer in line 5.

Excerpt 2-5 Verbal expression of enumeration in the interaction*

1. I: And uh how many hours or minutes did you watch television ? So just..
2. R: Oh well yes. Think Tank that started at seven thirty and then the news until twenty past eight and then I got uh guests at the door...
3. I: Yeah so uh...
4. R: Then I turned it off
5. I: ...about an hour.
6. R: Yes
7. I: Okay

*This Q-A sequence was taken from the Television Survey that is described in Chapter 5.

Alternatively, respondents may use generic information to form their judgment. For example, respondents may retrieve information about the rate of occurrence of events or behaviors, without recalling specific instances ('retrieved rate'). This strategy is likely to take much less time than enumeration. Therefore, respondents are less likely to express their thoughts. However, when they do express them, this may give indications about the adequacy of their strategy. As Houtkoop-Steenstra (2000) describes, respondents often add hedge expressions such as "I guess" or "probably" to their answers. From a validation study by Draisma and Dijkstra (2004) it appeared that such linguistic indicators of uncertainty occurred more frequently when responses appeared to be incorrect than when they were correct.

2.3.3 Formatting the response

When respondents have completed their mental judgment of an answer, they have to formulate it according to the response format available to them in the interview. Tourangeau et al. (2000) distinguish two processes within this step. The first is mapping the answer onto the appropriate scale or response options. The second, "editing" the response, entails that respondents adapt their answer to criteria such as consistency, social desirability, intrusiveness, or politeness. As editing is concerned with respondent's self-presentation towards the interviewer or third parties, it is less likely to be verbally expressed in the interaction, and therefore will not be discussed here.

The response format appears to have important consequences for the way respondents map their answer, and as a result may create interactional problems. Tourangeau et al. (2000) discuss three different types of very commonly used response formats: (1) Closed-ended items in which the response alternatives consist of an ordered set or a rating scale, such as agree-disagree scales and frequency ranges (e.g., 'once a week or less' 'twice a week' 'three times a week' 'more often'), (2) Closed-ended items in which the response alternatives comprise an unordered list (e.g., 'Democratic', 'Republican' or 'Independent'), and (3) Open-ended items in which respondents are required to give a numerical response.

The latter option is not an open-ended question in the sense that respondents have absolute freedom in formatting their response. The format implies an answer that is formatted as an exact number. The only difference with the closed-ended question of type (1), is that

interviewers (or questionnaires) do not mention the alternatives. However, the required format is indicated (e.g., 'how many days a week do you do X', 'what is the number of alcoholic units' etc.) or the end-points on the scale, implying all options in between them, are mentioned (e.g., 'a grade between 1 and 10'). Therefore, this type of question could better be referred to as items with implicit alternatives.

The freedom of formatting with implicit alternative items is problematic for several reasons. Respondents may give too general answers (e.g., ranges or approximations, i.e., mismatch answers), and also produce typical answers. Such typical answers are heaped at for instance multiples of 5, or in case of day-estimations concentrate on 'calendar prototypes' (e.g., providing answers heaping at 7 days, 30 days, or 365 days).

As Houtkoop-Steenstra (2000) notes, approximate numbers are in some cases culturally or language-specific. For example, when Dutch respondents give approximations, they often reply with 'about ten', whereas English speaking respondents often reply with ' a dozen', although there is reason to assume that both types of respondents may have the same approximation in mind. Nevertheless, interviewers easily accept 'about ten' to be a specific score of '10', and 'a dozen' to be a specific score of '12'. Consider the Q-A sequence in Excerpt 2-6. Initially, the respondent tries to formulate an answer at multiples of 5 (line 3), but after a suggestion of the interviewer (line 5), she adapts (line 8) this suggestion according to her idea of peculiarity of the number thirteen (line 6).

Excerpt 2-6 Problems with formatting the response in the interaction*

1.	I: What is the number of cups water, coffee, tea and other non-alcoholic drinks that you usually use on a day?
2.	R: Well that is different
3.	R: So between 10 and 15 cups
4.	I: Yes, and if you give an estimation, somewhere in the middle?
5.	I: Thirteen cups?
6.	R:Well thirteen that is such a strange number
7.	I: Yeah, that's true
8.	R:Make it fourteen

*This Q-A sequence was taken from the Health Issues Survey, which is described in chapter 7.

In case of closed-ended questions with an *unordered* set of alternatives, each of which are communicated to the respondent, other biases may occur. Krosnick (1999) argues that primacy and recency biases may be indicative of weak 'satisficing' behavior. Respondents are likely to select the first reasonable response they come across, because the more alternatives they have considered, the less motivation or ability remains available to fully consider subsequent alternatives. When interviewers quickly read the alternatives, it is likely that the last alternative is most actively present in the working memory of the respondent. Therefore, respondents are likely to choose this alternative, thus yielding a recency bias (Krosnick 1999). However, a primacy bias may also occur due to the fact that unordered alternatives often do not project a transition relevance place (see section 2.2.9), which may

cause respondents to interrupt the interviewer with an answer before she has read all alternatives, thus yielding a primacy bias.

In case of response alternatives with an *ordered* set of alternatives, respondents may draw inferences from the response scale, assuming that a scale is based upon the researcher's knowledge about the dispersal of the behavior in the population. This takes them to assume that the average or normal frequency is represented by the middle response alternative, and that extreme values are represented by the endpoints of the scale. This assumption is based upon the Gricean principle of cooperation, which triggers respondents to assume that everything the researcher communicates is meaningful and informative (Schwarz, 1996, see also section 2.2.7).

Vague quantifiers (such as 'not too often', 'pretty often' 'always') are also not an ideal option, as they specify relative positions, but do not reflect an absolute frequency. Although all respondents will agree that 'pretty often' is more than 'not too often', and is less than 'always', it is also clear that the respondent's evaluations of the exact differences between the categories will differ, not only between respondents, but also between different contexts for the same respondent (Schaeffer and Presser 2003, Tourangeau et al. 2000). Therefore, respondents are likely to request for clarification, when they are asked questions with vague quantifiers.

2.3.4 Additions to the model of survey response

As Krosnick (1999) notes, due to the large amount of cognitive processing that is required to thoroughly go through all four steps of Tourangeau et al. (2000) model, it is unrealistic to assume that respondents take this effort to generate an optimal answer. Response behavior that is likely to occur is 'weak satisficing', which involves executing all steps, but less carefully than in case of 'optimizing' strategies. Respondents may also skip steps, such as retrieval and judgment. Those 'strong satisficers' base their answer upon cues in the question "pointing at a response that can be easily selected and defended if necessary" (Krosnick, 1999, p. 548).

Cannell, Miller and Oksenberg's model (1981), resembles Tourangeau's et al. four step model, but includes the notion of different routes, one based upon careful processing of questions, the other based on superficial cues. As Tourangeau et al. note, Cannell's et al. notion of different routes, has a parallel with the Petty and Caccioppo's elaboration likelihood model (Petty and Cacioppo 1981). According to this model, receivers of a message can follow two different routes of information processing. The first route is the central route. This entails a rather systematic processing of information, involving careful consideration of issues related to the message. The alternative route, i.e., the peripheral route, does *not* involve thorough processing; only simple cues are used to process information of a message.

In case of questions, it is reasonable to assume that a more thoroughly processed question wording yields answers of better quality. Therefore, it is interesting to examine which variables may affect the elaboration of messages, i.e., to find out why receivers follow the central or the peripheral route of information processing. Petty and Cacioppo found that

(among other findings) the receivers' ability to elaborate a message may be decreased by distraction (which is in accordance with Holbrook et al.'s findings, see section 2.2.16).

Petty, Rennier and Cacioppo (1987) argue that "just as people are often unmotivated to think about persuasive communications, so too they may sometimes be unwilling to devote their limited cognitive resources to thinking about the issues raised in opinion surveys" (Petty, Rennier and Cacioppo 1987, p. 485).

As Tourangeau et al. (2000) conclude, Cannell's et al. (1981) model including two routes concentrates on the respondent's decisions to answer accurately, whereas Tourangeau et al. favor an emphasis on a model of four steps, without necessarily suggesting that respondents perform all four steps when they answer a question. The distinction between careful and superficial processing is likely to become fuzzy. For example, previous questions may automatically affect retrieval of subsequent similar questions, regardless of the fact that respondents try to process questions carefully. Therefore they conclude that the two routes can best be viewed as "two extremes on a continuum of processes that vary in the depth and the quality of thought that respondents give to their answers" (Tourangeau et al., 2000, p.17).

2.4 Models accounting for the interaction and cognitive processes in the survey interview

A striking characteristic of most cognitive models of survey response is that the interviewer is not included as a factor. Although, according to Tourangeau et al. (2000) this is a matter of 'emphasis', we think that when a more complete cognitive model of survey response in a survey interview is presented, it is better to include the interviewer. In this section we will not only describe models that account for the cognitive processes of the interviewer and respondent, but also attempt to describe models that include the interaction between the interviewer and the respondent.

2.4.1 Models accounting for cognitive processes of the interviewer

Sander et al. (1992) mention three models that account for the mental processes taking place by the interviewer; 'The interviewer model of question generation', 'The interviewer model of question clarification' and 'The interviewer-respondent interaction' model. When added to 'The respondent model of question answering' (i.e., Tourangeau's model of survey response) these four models illustrate the "major information processes and behaviors of the interviewer and respondent" (Sander et al. 1992, p. 818). Unfortunately Sander et al. do not describe the interviewer-respondent interaction model. This shortcoming is not surprising as, like we pointed out in section 2.2, the interaction between the interviewer and respondent may be very complex to present in a model.

2.4.2 Adaptation of Sander et al.'s Interviewer Model of Question Generation

Sander et al.'s 'interviewer model of question generation' basically constitutes two pathways of processes that lead to either verbatim or non-verbatim question reading. We have slightly adapted and elaborated this model in Figure 2-2. The model, like Sander et al.'s, concerns

completion of *one* question within a survey. The upper and right part of the figure comprises a summary of Sander et al.'s model. According to this model, the number of times an interviewer had conducted the same interview and has posed the same questions previously (a) will influence the availability of question wordings in the interviewer's memory (b). When the question wording is not available in the interviewer's memory, she will have to read the question text (c), and subsequently, depending on her commitment to verbatim reading, encode the question words (p), or interpret the question's meaning (n) to pose the question (r). When the question wording is available in the interviewer's memory, interviewers may reconstruct the question's meaning (o) or even completely reproduce the question from memory (q), by only glancing (d) at some question words, the question number, or the specific page or screen within the questionnaire.

As we discussed in section 2.2, interviewers may have several reasons to not read questions exactly as worded, which is visualized in the 'commitment to verbatim reading' box. As Sander et al. also argue (but not visually present in their model), a large part of the questions that are not presented verbatim may, besides reading and speech errors, be accounted for by more or less deliberate paraphrases of the original question. Interviewers may paraphrase questions when the original question is lengthy, awkwardly structured or just difficult to read, as judged for themselves (i) and for respondents (e). Moreover, interviewers may anticipate difficulties from respondents in understanding questions (f, see section 2.2.13), use paraphrases to establish rapport to avoid asking questions offensively (g, see section 2.2.8) or account for information already provided (h, see section 2.2.12). Furthermore, repeatedly asking the same questions over and over again across different interviews (j) may be boring for interviewers (and for respondents, see section 2.2.3) which may cause interviewers to vary their question wording.

Nevertheless, it is important not to overestimate the occurrence of variance in question reading. Although the percentage of questions that are read exactly as worded may depend on the specific survey, and the definition a researcher holds for coding of 'exactly as worded', a percentage above 80% is often found. Interviewers may be lent a hand to read questions exactly as worded when question wordings are carefully pretested (k), and when interviewers are well trained (l) and provided with clearly presented questions (m), e.g., using large font and a lay-out that clearly distinguishes questions from instruction texts.

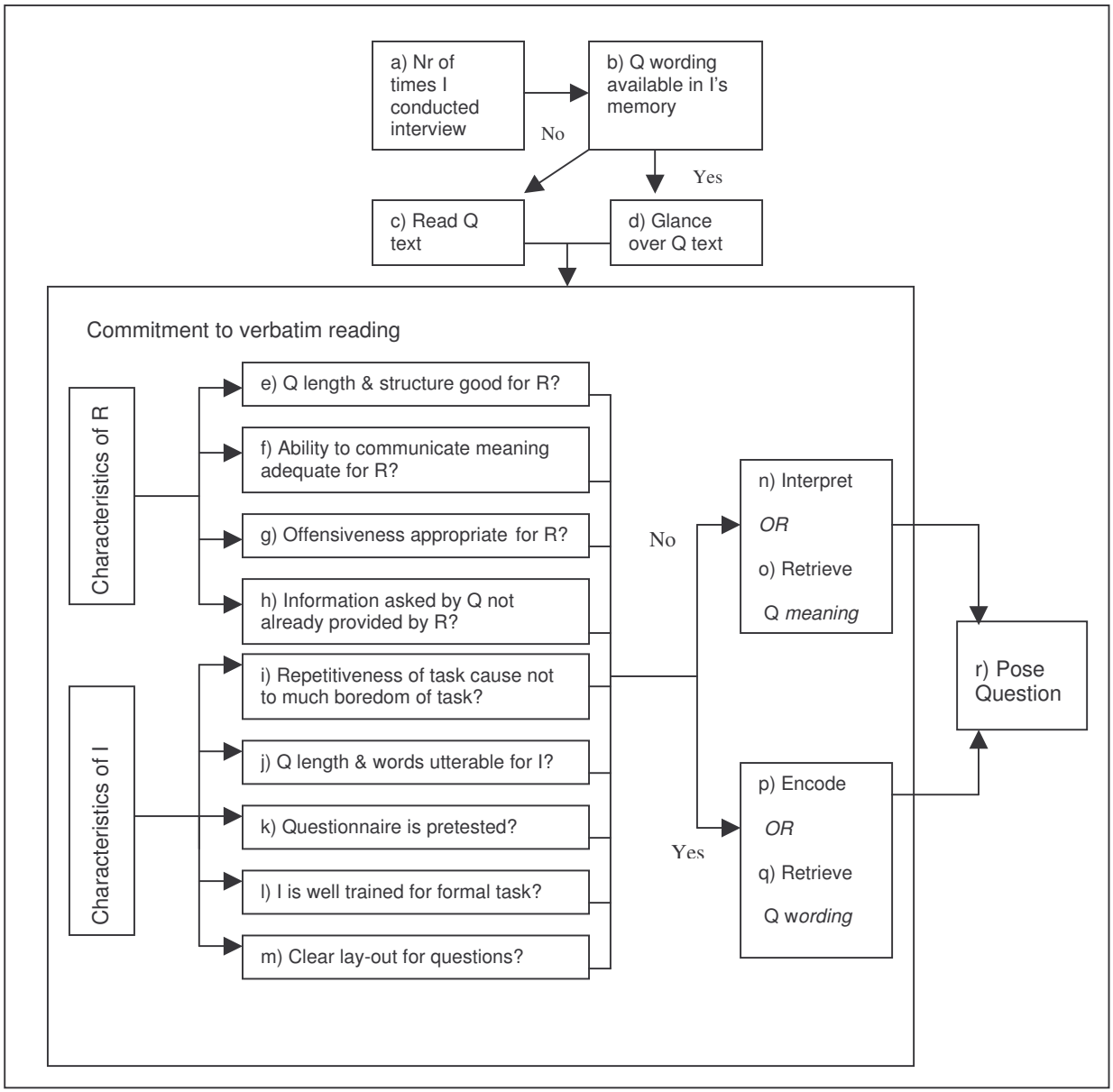

Figure 2-2 Model of Interviewer's Question Formulation (Partly based on Sander et al.'s 'The Interviewer Model of Question Generation').

2.4.3 Adaptation of the Respondent's model of Question Answering and Sander et al.'s Interviewer Model of Question Clarification.

In Figure 2-3 we summarize the cognitive processes we described in this chapter (except for 'comprehension' that was already summarized in figure 2-1). The left part of the figure (i.e., boxes a-d) is based upon the Tourangeau et al. respondent model of question answering, but the model also includes the interactional processes described in section 2.2. It is the first attempt to present an interviewer-respondent interaction model.

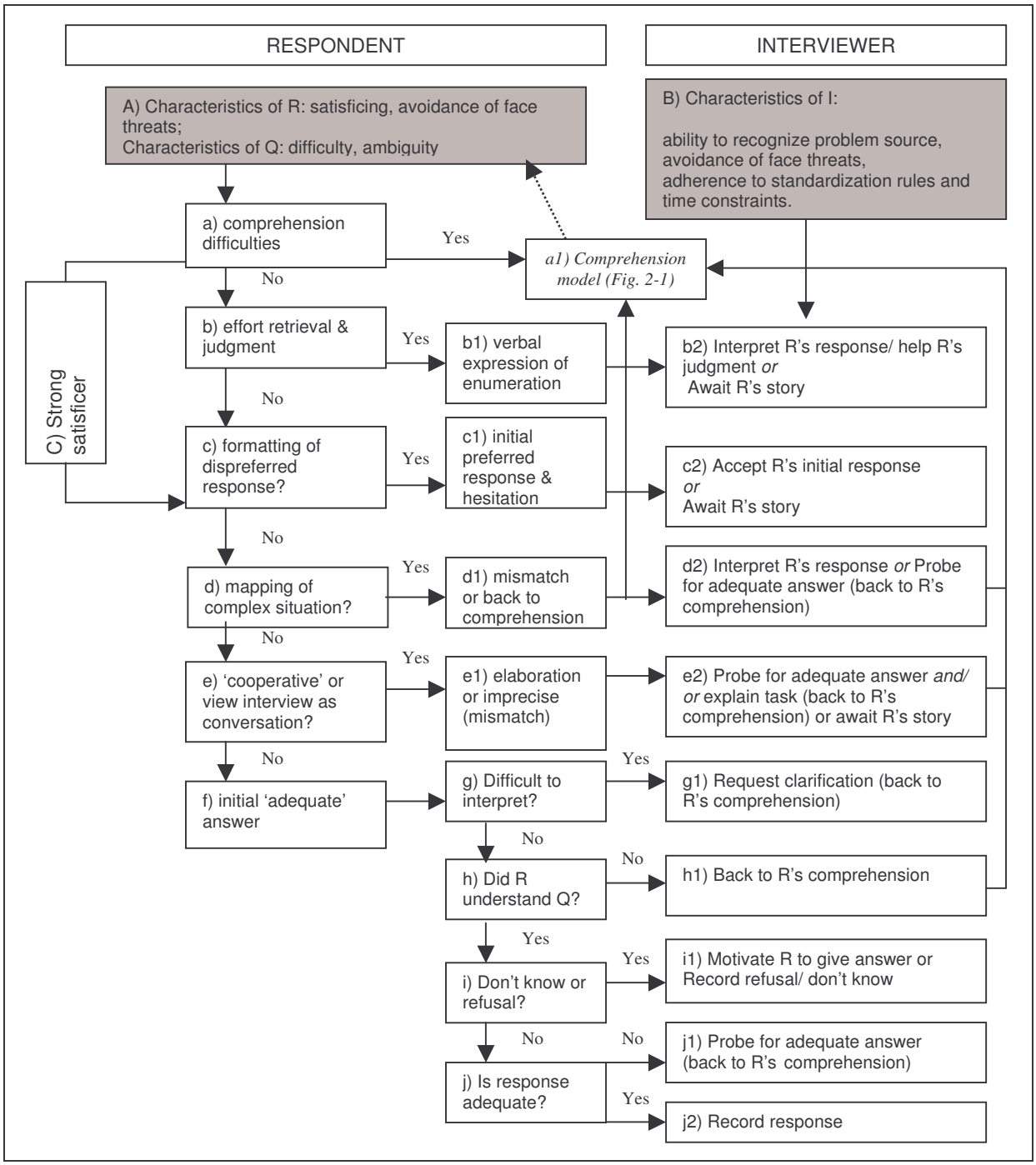

Figure 2-3 Model of respondent-interviewer interaction in the retrieval, judgment and formatting phases

After the respondent and interviewer have dealt with any comprehension difficulties (a, see Figure 2-1), a respondent may verbally express retrieval and judgment. The effort a respondent may put into retrieval and judgment (b) may cause a respondent to verbally express this retrieval and judgment phase by means of enumerations (b1). The interviewer's reaction depends on her patience, or the assumption that the respondent will be in no need for help with his enumerations, and the interviewer's ability to correctly recognize the problem

source (B). Thus, the respondent's verbal expressions may subsequently trigger the interviewer to offer some help, or even infer the respondent's answer, but the interviewer may also await the respondent's full story of enumerations (b2).

With respect to response formatting, several problems may occur. The respondent may be reluctant to give a dispreferred response (c). According to the preference for agreement (see section 2.2.11) respondents may start with an initial preferred response (c1). The interviewer may quickly accept this preferred response, leaving no room for the respondent to adapt his answer to the dispreferred response or await the respondent's story (c2). The respondent may also be faced with a complex situation that is difficult to map on the response alternatives (d). This may lead the respondent and interviewer back to the comprehension phase, but the respondent may also attempt to formulate an answer that comes close to the response alternatives (i.e., a mismatch answer, d1). The interviewer may attempt to solve the respondent's problem with her own interpretation of the mismatch answer, but she may also probe for an adequate answer (d2). Furthermore, a problem with formatting may occur when the respondent views the survey interview as an ordinary conversation (e), during which elaborations are allowed and precise answers are not obligatory, which may result in elaborations or a mismatch answer (e1). Depending on the interviewer's tendency to avoid face threatening actions and her dependency on time constraints, the interviewer may probe for an adequate answer, explain the respondent's task, or just let the respondent talk (e2). This not only lengthens the interaction but may also distract the interviewer and respondent from their tasks, and may have consequences for the interpretation of the response (see section 2.2.15).

The bottom part of the figure represents a summary of Sander et al.'s 'interviewer model of question clarification'. However, this model actually illustrates more than clarification alone, as it constitutes more general processes involved in the communication of the respondent's answer. After a respondent's initial 'adequate' answer (f), the interviewer interprets the meaning of this response (g), subsequently may ask for clarification of the response (g1), and decides whether the respondent provided an answer that indicates an incorrect understanding of the question (h). If the latter is the case, the interviewer will repeat the question or provide clarification, leading the respondent back to comprehension. Furthermore, the interviewer will establish whether the answer comprises a 'don't know' or refusal (i). If this is the case, the interviewer may motivate the respondent to think about giving a substantial answer, or just record 'don't know' or 'refusal'. If the answer is a substantial one, finally, the interviewer will establish the adequacy of the answer (j). If the respondent's answer is not adequate, the interviewer will probe (j1) and else scores the answer (j2).

As we illustrated in section 2.2, the interviewer-respondent interaction is rather complex. It appears that a lot of details in this interaction may yet remain to be detected. For example, the model does not specify exactly when interviewers are able to recognize problem sources, when they avoid face threats, and when they adhere to standardization rules or time constraints. Furthermore, specific problems may lead to several different behaviors and these

behaviors may have several interactional causes. The model in Figure 2-3 also does not include details of the various variables that influence the cognitive processes.

2.5 Conclusion

In this chapter we have given a review of the models of interviewer and respondent tasks, cognitive theories, and theories about the difference between ordinary conversations and standardized survey interviews. This review not only indicates what aspects of the survey interview are problematic and are worthwhile studying, it also provides us with ideas about which behaviors of interviewers and respondents can be relevant to detect problems in interactions.

As verbal behaviors as such will provide a wealth of information on their own, in this thesis we will exclude non-verbal behaviors. Although cognitive theories suggest that response times may provide a lot of information about the cognitive capacity of respondents, the difficulty or even the sensitivity of question topics, we assume that verbal behaviors can also be used as indicators of cognitive problems. Furthermore, we will limit our focus on verbal behavior to the pragmatic relevance of utterances. We will not include very detailed linguistic aspects of verbal behaviors, such as the tense in which questions and answers are formulated, but we will determine the adequacy of questions and answers and other task-related utterances.

In order to answer our research questions, we need to know more about systematic patterns in interactions, what problematic deviations typically occur, and what the causes are of these deviations. Hence we should systematically study the interaction in order to test and elaborate the models of interviewer and respondent interaction as presented in this chapter. A number of methods are available, and will be described in the next chapter.

3 Methods of Behavior Coding of Survey Interviews[4]

3.1 Introduction

In the previous chapter we presented models of interviewer-respondent interaction. The models showed that this interaction can be very complex. Thus, in order to describe the interaction and study relations, we need an efficient method to analyze these data. Behavior coding comprises a systematic coding of interviewer and/or respondent behaviors in survey interviews. The process of questioning and answering in the survey interview takes place in so-called question-answer sequences (Q-A sequences), which comprise all utterances of interviewer and respondent that belong to a survey question.

Both the interviewer and the respondent can cause deviations from the paradigmatic Q-A sequence (see section 1.3). In a broad sense, behavior coding is intended to discover departures from the paradigmatic sequence, and to discover how these departures relate to data quality on the one hand, and characteristics of interviewer, respondent, or questionnaire design on the other hand. Paradigmatic sequences usually make up the largest part of Q-A sequences in an interview, but may vary from for example 35% to 95% of the Q-A sequences for different questions within the same survey (Van der Zouwen and Dijkstra 1998).

In 1968, Cannell, Fowler and Marquis devised the first, fairly simple scheme to code behavior in the standardized survey interview. Next, coding schemes generally became more and more sophisticated as well as more complex, as with each subsequent coding scheme and its application to actual data, more and more became known about the interaction between interviewer and respondent. In addition, the development of more sophisticated coding schemes was stimulated because technical devices became available. Especially the availability of the tape recorder may explain the increase in the number of codes that were included in the coding scheme. The scheme of Cannell, Fowler and Marquis (1968) includes only 12 different codes, and did not rely on the availability of tape recorders. In a subsequent study, Marquis and Cannell (1969) did use tape recordings, and described a far more detailed coding scheme, consisting of 47 different codes.

The increase in number of codes that could be included in coding schemes is even more stimulated by a second technical device that could be used for behavior coding. This device was the computer. A program like the Sequence Viewer program (Dijkstra 1999; Dijkstra 2002) enabled the coder to quickly and reliably enter a lot of different codes, whereas the coding could also be carried out semi-automatically, based on the transcripts. The text analysis options in this program enable automatic coding of all paradigmatic Q-A sequences. However, the increased feasibility to enter large amounts of data was not the only benefit of the use of computers. The possibility to *analyze* a large number of codes and large data sets was another major advantage of using computers. Because of that capacity, it became worthwhile to invest in the time-consuming process of transcribing and coding interviews in a detailed way. For example, Loosveldt (1985) describes that for the analysis of the 11.331

[4] This chapter is also forthcoming as: Ongena, Y.P. and W. Dijkstra (forthcoming) "Methods of Behavior Coding of Survey Interviews." *Journal of Official Statistics.*

actions that were coded, special programs were written. The Sequence Viewer program also allows researchers to perform a large number of different, more and more sophisticated analyses (Dijkstra 2002).

The number of different categories included is probably the most obvious difference between coding schemes. The number of categories varies from two values (Edwards et al. 2002) to around two hundred different code combinations in an average dataset (Dijkstra 1999).

It is beyond the scope of this chapter to give a full account of all codes used in the 48 coding schemes that were studied, but we will discuss some common distinctions. We found 134 different categories for interviewer behavior, 78 different categories for respondent behavior, and 14 different categories for behavior of third parties. In Table 3-1 examples are listed of typical behavioral codes, which are used in at least 12 (i.e., 25%) of the 48 coding schemes evaluated in this chapter. The table also shows for all codes the number of schemes that include the code, and the range in percentage of occurrence of the behavior in Q-A sequences as reported in the studies that used the code.

Table 3-1 Most common codes included in coding schemes and average reported frequency of occurrence in Q-A sequences

Interviewer Behavior codes	Nr of coding schemes	Range in % of occurrence	*Respondent Behavior codes*	Nr of coding schemes	Range in % of occurrence
Question reading exactly as scripted	26	28-97%	Adequate answer	25	75-95%
Question read with minor change	21	1-32%	Inadequate answer	21	2-27%
Question read with major change	35	0-25%	Don't know answer	17	1-6%
Question skipped/not verified	16	0-22%	Refusal to answer	21	0-1%
Non-directive probe in interviewer's words	23	5-80%	Request for clarification	18	0-23%
Suggestive probe	15	0-33%	Interruption	18	0-36%
			Qualified answer	14	2-20%

Cannell and Oksenberg (1988) indicate that the kind of code categories that are included in a coding scheme depend upon the research objective. However, this appears to be only partially true; irrespective of the focus of the scheme, most schemes include codes for interviewer's question reading.

For behavior coding as a proper diagnostic tool, it is important that all relevant behaviors are included in the coding schemes. It may not always be possible to determine in advance what those relevant behaviors are, and therefore the development of a behavior coding scheme can be considered as an iterative process.

As Table 3-2 shows, behavior coding is typically related to variables in the data collection procedures (such as question wording and interviewer styles), and can be

implemented in different phases of survey data collection. Results of behavior coding implemented prior to or during actual data collection, can be used to adapt data collection procedures. Behavior coding data can also be used as dependent variable in experiments (e.g., comparing question wordings or differently trained interviewers). It can also be used as independent variables in studies that aim to detect relations between problematic behaviors and the validity and reliability of scores obtained (Belli and Lepkowski 1996; Dijkstra and Ongena forthcoming; Dykema et al. 1997). In this thesis, behavior coding is only applied after actual data collection (i.e., the last four rows of Table 3-2). We used behavior coding to explore causes and effects of behaviors (chapter 5), to evaluate data quality, and we use behavior coding data as a dependent variable (chapter 6). We used it also to check experimental manipulations, and as a dependent variable of experimentally manipulated data collection (chapter 7).

Table 3-2 Possible implementations of behavior coding

Goal	Phase of study
Pretest of questionnaire, interview mode etc.	*Prior* to actual data collection
Monitor interviewers.	*During* actual data collection
Evaluate data quality, functioning of interviewers and respondents, effectiveness of revisions, explain biases in response distributions	*After* actual data collection
Explore causes and effects of behaviors	*After* actual data collection
Checking experimental manipulations	*After experimentally* manipulated data collection
Use behavior coding as a *dependent* variable	*After experimentally* manipulated data collection

In this chapter an exhaustive overview is given of all applications of behavior coding, comparing characteristics of 48 coding schemes,[5] presented in manuals, conference proceedings, articles, etc. Advantages and disadvantages of different strategies and procedures will be given. Finally, we give recommendations about the types of coding schemes and procedures that are most appropriate in specific situations.

3.2 Coding strategies

Some fundamental decisions in the design of a coding scheme have consequences for the applicability of the scheme. These decisions concern the unit of coding, whether full or selective coding is applied, and whether and how sequence information will be preserved.

[5] In this comparison of coding schemes only first published articles of coding schemes are included. Coding schemes of the same author(s) that underwent important changes (either in the codes included or in the coding procedures) are treated as separate cases of coding schemes.

3.2.1 Units of coding

A common strategy in coding social behavior is to use time-intervals as a unit to assign codes to (Bakeman and Gottman 1997). However, survey researchers are much more interested in particular types of utterances of interviewers and respondents, irrespective of their length. In none of the schemes examined, time intervals were used. Behavior coding appears to take place at one of four different levels; i.e., utterances, exchange levels, Q-A sequences or entire interviews.

Coding at the utterance-level

A strategy that is especially useful in case of interaction analysis, is coding at the level of the utterance. Each utterance can get one code, but not more than one code. It is not possible to code behavior that did not take place, e.g., the absence of an adequate answer. However, if full coding is applied (see section 3.2.2), and/or sequence information is preserved it is possible to infer the absence of certain behaviors from the coded utterances within a Q-A sequence.

To code the utterances of a Q-A sequence, the sequence should be separated into meaningful parts. The turn is too rough as a segmentation procedure, because it may consist of multiple 'turn-constructional units' (TCU's, see section 2.2.9). When coders try to determine the appropriate codes, most problems occur as soon as utterances are not adequately segmented into separate TCU's. Multiple types of behaviors can be performed within a turn. As a result, multiple codes may be applicable to one turn, which creates a problem for the coder.

Therefore, it is important that the utterances in Q-A sequences are carefully segmented into TCU's. According to pragmatic completeness, a TCU is complete when the utterance is recognizable as an independent informative and functional unit. Pragmatic completeness is assessed by means of sequence reasoning: i.e., the sequential position of an utterance as part of utterances that are functionally related (Mazeland 2003). Segmenting the utterances consists of judging the pragmatic completeness of utterances, whereas coding the utterances consists of applying a pragmatic description to an utterance.

Coding at the exchange level

It is possible to code at a level that is intermediate between the utterance and the Q-A sequence level; this intermediate level is often referred to as the exchange level. An exchange can be considered as an adjacency pair of a question and an answer. Typically, the first two exchanges are coded; i.e., (1) the exchange of initial question reading and an initial response, and (2) the exchange of a prompt by the interviewer and a possible second answer by the respondent. The coder must ignore insignificant behaviors that may occur in between (i.e., neutral acknowledgement tokens, silences, laughter) and ignore anything after the second answer. Morton-Williams (1979) was the first to use this kind of coding. Such a coding strategy is selective with respect to the part of the Q-A sequence that is coded, but it still

enables preservation of sequential information, which is not possible in case of coding at the Q-A sequence level.

Coding at the Q-A sequence-level

Assigning a code to the whole Q-A sequence, may comprise judging whether or not a specific type of behavior takes place in the Q-A sequence, or whether the Q-A sequence is paradigmatic or problematic. The division of units to be coded is in this case more straightforward: a Q-A sequence starts as soon as the interviewer starts reading a question, and ends as soon as the next question is posed. However, it is of course possible that, whereas the interviewer has posed a next question, the respondent elaborates his answer to the previous question. Such behaviors may be easily overlooked, or assigned to the wrong Q-A sequence, especially when coding does not take place from transcripts (see section 3.3.1).

As compared to coding at the utterance or exchange level, coding at the Q-A sequence level is more sensitive to errors of omission. According to Cannell, Lawson and Hausser (1975), disagreements in coding of entire Q-A sequences often do not concern which particular code should be used for a behavior, but rather upon whether or not a particular behavior should be coded at all.

Coding at the interview-level

A final unit is the whole interview, e.g., if the whole interview is assigned some evaluative code. Carton (1999) for example added codes to characterize the whole interview with respect to specific interviewer behaviors such as giving instructions, asking questions and probing, and general evaluations such as the orientation towards the respondent and the atmosphere during the interview. In the comparison of behavior coding schemes we did not include schemes that only use coding at the level of the interview (e.g., Brick et al. 1997a; Mathiowetz 1999).

3.2.2 Full or selective coding

A fundamental difference between coding schemes, is that coding can be applied to all utterances (‘full coding’) or to a selection of utterances or behaviors that are considered as important or relevant for the specific research question (‘selective coding’). Selective coding schemes are essentially developed from a practical point of view: it is determined in advance what behaviors are diagnostic of problems that the researcher wishes to detect. For example, if one studies general interviewer performance, only interviewer behaviors are coded.

A full coding scheme is often used when the researcher’s goal is to explore the interaction. With full coding data it is possible to reconstruct more or less what occurred in an interview. Full coding must take place at the utterance level, as it requires assigning a relevant category to each utterance, whereas selective coding may take place at the QA-sequence level but also at the utterance or exchange level. In the latter two cases, it is possible to preserve sequential information at the exchange level. For example, in Cannell’s et al. (1975) coding scheme, only interviewer behaviors were coded (therefore constituting a

selective coding scheme at the utterance level). Nevertheless, they instructed the coders to code in the order of occurrence, and all respondent utterances in between the interviewer's utterances were represented by vertical lines.

The combination of the three levels of coding and application of full or selective coding yields six possibilities, of which only four are relevant, because full coding can only take place at the utterance level. Hence, we can distinguish four coding strategies; full coding of utterances, selective coding of utterances, coding at the exchange level and coding of whole Q-A sequences. These strategies have different consequences for the possibility of preservation of sequential information, as shown in Table 3-3.

Table 3-3 Overview of coding strategies and possibilities of preserving sequential information

Strategy	Unit of coding	Sequential information applicable
Full coding	Utterance	++
Selective coding	Utterance	+
	Exchange	+
	Q-A sequence	-

In Table 3-4 advantages and disadvantages of three coding strategies are shown. With respect to *quick results* and *practical feasibility* coding at the Q-A sequence level is rated highest. This strategy makes quick results possible, without the use of specialized software. For instance, coders may only have to note inadequate readings of questions or requests for clarifications from respondents.

Table 3-4: Overview of advantages and disadvantages of coding strategy

	Selective coding: whole Q-A sequence	Selective coding: utterances or exchanges	Full coding: utterances
Quick results	++	+	-
Practical feasibility	+	-	- -
Simplicity	+	+	-
Absent behavior	+	-	+
Amount of information	-	+	++
Sequence information	- -	+	++
Identification of paradigmatic sequence	+	-	++

Selective coding at the exchange or utterance level takes a medium position on practical feasibility. Full coding is by far the most tedious kind of coding. In order to apply full coding, it is important to have software available that facilitates the transcribing, coding and analyzing of the data. Without such software, full coding with sequential information is hardly feasible.

As Smit (1995) notes, it is important that the number of codes included in a coding scheme is manageable; with too detailed coding schemes it will often be problematic to employ clear methods of analysis. Moreover, with a complex coding scheme the coding process will be more error-prone and time-consuming. The *simplicity* of the coding scheme is highest in case of selective coding at the Q-A sequence or exchange level. For full coding a detailed, and consequently complex coding scheme is necessary to meaningfully characterize all the various behaviors that can occur during an interview. However, several options are available to enhance the simplicity of the scheme (see section 3.3.4).

Whole Q-A sequences can easily be coded according to the *absence* of relevant behavior. In case of full coding, absence of behavior may be inferred from analysis of complete Q-A sequences.

The *amount of information* will usually be lowest in case of coding at the Q-A sequence level, and hence potentially important behavior may easily be overlooked. If sequential information at the exchange level is preserved, even fairly simple coding schemes yield a lot of extra information, although here too, significant behaviors may easily be overlooked. Most information, also about the *sequence* of behaviors, is available in case of full coding.

Full coding provides a researcher with information about any *deviation from a paradigmatic sequence*. In case of coding at the Q-A sequence level, it is possible to include codes that evaluate the Q-A sequence as a whole. In case of selective coding of utterances or exchanges, it is difficult to obtain information of all deviations from paradigmatic sequences. In all cases of selective coding, it is possible that deviations that are not coded are more indicative of problems than the coded ones.

3.2.3 Type of analysis

Two main types of quantitative analysis of behavior coding data can be distinguished, i.e., frequency analysis and sequence analysis. Furthermore, quantitative analyses may be supported by qualitative analyses of the actual interactions, provided that transcripts are available.

Frequency analysis

Frequency analysis essentially concerns counting the occurrence of particular types of interviewer and respondent behavior. The frequency of occurrence of specific behaviors may be related to other factors, like interviewer or question characteristics, or response distributions. For example, Edwards et al. (2004) compared frequencies of interviewer and respondent behaviors across interviews of the same questionnaire in different languages. One of the findings was that respondents appeared to behave differently when they are being interviewed in their first language (i.e., interrupting the interviewer and giving extraneous comments more often) than in a second language.

Furthermore, frequency analysis can be used in experimental designs that compare different question wordings, different interviewing styles, or other manipulations of data collection procedures in survey interviews. For example, one can establish the effects of

different question wordings on the occurrence of inadequate answers. In that case it is important to verify that interviewers have read the questions exactly as worded.

Frequency analyses can be supplemented with analyses of variance or log-linear analyses at the Q-A sequence level (i.e., comparing question, interviewer or respondent variables with average number or odds ratios of problematic behaviors occurring in the Q-A sequences).

Sequence analysis

Sequence analysis allows studying dependencies between different types of behavior, in particular the relation between subsequent interviewer and respondent behaviors. In case of selective coding schemes, sequence analysis is rather limited; it is possible to distinguish initial from secondary responses, and initial question asking from follow-up probing, but not for example what kind of non-problematic behaviors may have occurred in between questions and answers.

In order to be able to interpret results of sequence analysis correctly, it is important that the assignment of codes is independent from codes that precede or follow the behavior to be coded. In some cases it is hardly avoidable that coding a particular behavior depends on previous utterances. A code for 'interviewer repeats respondent's answer' is likely to be preceded by an answer of the respondent, but it is hardly possible not to take the preceding utterance into account. Nonetheless, assigning a particular code should never depend on subsequent behavior, to prevent that relations between behavior and subsequent behavior are artificial.

Data that are generated through full coding schemes enable analyses by means of a *tree* representation of the structure of interviewer respondent interaction. Brenner (1982) was the first researcher to present such a tree analysis. A tree may represent the consequences of a particular action of either interviewer or respondent. From other analyses it is possible to analyze the causes of particular actions of interviewer or respondent. For example, with the lag-sequential analysis that Smit (1995) describes, it is possible to determine which parts of subsequent behaviors in a Q-A sequence occur below or above chance.

Supplementary analyses

Behavior coding studies concerning the frequency of occurrence of behaviors very often only offer data from tables but do not uncover sources of problematic behaviors. It often remains unclear, even in case of sequential analysis, how events in the interaction can have certain causes or effects; i.e., what actually happened in the interaction.

One way to learn more about this, is to use code frequencies as input for discussions with interviewers or coders (i.e., debriefing, see Oksenberg, Cannell and Kalton 1991). Using coders for debriefing is useful because coders are not personally involved in the interviews, and having listened to tape recordings, have full access to relevant information of the interactions (DeMaio et al. 1993). Notes of coders are often used to diagnose the source and the seriousness of the problems (e.g., Dykema et al. 1997, Schaeffer and Dykema 2004).

Such notes may specify a major change in question reading, with abbreviations to indicate the nature of the change (addition, deletion or other) and the indications of the specific words that were added or deleted (Schaeffer and Dykema 2004).

However, the actual conversations on tape could be even more useful. It is quite possible that coders do not notice all aspects that are worthwhile inspecting with more detail. Furthermore, transcripts can easily illustrate findings. Finally, other sources of information can be used, such as answer distributions, response latencies (see Draisma and Dijkstra 2004), and details of the date, time and location of the interviews.

3.3 Practical considerations in coding procedures

The coding procedure is an important feature for the usability and reliability of a coding scheme. According to Cannell and Oksenberg (1988) it makes little difference whether the observation mode comprises face-to-face or telephone interviews, and whether live coding or coding from tape recordings is used, because the techniques for coding behavior are the same. However, they ignored the procedure of using transcripts, which is hardly to be avoided in case of full coding, but an option in case of selective coding.

3.3.1 Live coding, coding from tape and using transcripts

Coding can be done during the interview ('live coding') or afterwards, by listening to tape-recorded interviews ('recorded coding') or by using transcripts of the tape-recorded interviews ('transcript coding'). The advantages and disadvantages of these three procedures are summarized in Table 3-5.

Table 3-5 Overview of advantages and disadvantages of different coding procedures

	Live coding	Live coding with tape as backup	Recorded tape coding	Recorded transcript coding
Cost	++	+	-	- -
Permission	+	-	-	-
Unobtrusive	-	-	+	+
Efficient planning	-	-	++	+
Reliability	-	-	+	++
Semi-automatic coding	-	-	-	+
Check of coder performance	-	+	+	++
Paralinguistics	-	-	+	-
Thorough analysis	- -	-	+	++

Costs, in terms of time and workload, are lowest for live coding and highest for recorded transcript coding. In only six studies some indication is given of the time involved in coding interviews (including transcribing or otherwise). This ranges from a time equal to the interview, in case of live coding, to about six times the duration of an interview, in case of transcript coding.

The advantage of live coding is of course that data are immediately available; it is finished concurrently with the interview. Coding from tape may be more *efficient* than live coding, because coders do not have to wait for an interview to occur (DeMaio et al. 1993). Furthermore, tape coding is a relatively quick method, because no transcripts are produced. However, the additional time that is needed for producing transcripts may be partly regained when complex Q-A sequences are coded. In that case, transcripts may help coders to see the complete Q-A sequence. With this information it is easier to determine what code is appropriate, and in case of doubt it is possible to just read again the utterances in the transcript, instead of rewinding the tape to search for the fragment.

In case of live coding, *permission* to record the interview is not necessary of course. However, live coding in case of personal interviews may be more *obtrusive* than coding from tape or transcripts, because a coder needs to be present during the interview.

Although live coding can be *reliable* (Esposito et al.1992), recorded coding will always enable better quality of coding, as coders have more time to decide on the most appropriate code, and can consult code descriptions. When coding takes place from transcripts, reliability can be even more improved. Transcript coding in fact comprises a coding procedure in three steps (transcription, segmentation of meaningful utterances, and coding, comprising assignment of meaning to utterances). The researcher may perform separate reliability checks for the latter two tasks (see Smit 1995), or even decide to assign the different tasks to independent transcribers and coders.

Whenever coding takes place live or directly from tape, it is likely that important, meaningful behaviors are ignored. It is important that coders have useful visual documents available that enable them to compare what they hear on tape with the exact question wordings and the interviewer's recordings. Completed questionnaires or responses that are copied onto blank questionnaires may be an alternative to transcripts (Cahalan et al. 1994). However, especially complex coding schemes will require transcripts to warrant reliable coding. As Dijkstra (1999) points out, coding from transcripts can be done *semi-automatically* for utterances that occur frequently.

Tape coding enables good possibilities for *checks of coder performance*, but transcript coding also enables more systematic checks. Determining inter-coder reliability in case of live coding is only possible by means of having multiple coders code simultaneously. However, a live-coded interview may be taped as well, to allow checking samples of the coding and to (re)code or correct complex parts in the interactions. In that case some advantages of both recorded and live coding are combined.

In some cases special attention must be paid to *paralinguistic* features of the utterances. A different tone and accent can for example change the meaning of an utterance. When just the written text is used for coding, errors might be made as a result of ignoring these features. It is therefore important to have sound files easily available when coding from transcripts.

Obviously, recorded coding as compared to live coding increases the options in the complexity of the coding scheme and thus makes more *thorough analysis* possible. But, as noted before, transcripts certainly will be helpful to illustrate or explain results from plain

analysis of the codes. When the interview is coded from tape, it will be less likely that effort will be invested to find the fragment that illustrates a certain result.

It appears that recorded tape coding is the most popular procedure, as in 31 of the 48 schemes this procedure was followed. The difference between live coding and recorded coding is clearly illustrated by the number of codes included in coding schemes. Schemes that are designed for live coding contain between 2 and 20 codes (median: 13 codes), whereas schemes designed for recorded coding contained 2 to 174 codes (median: 22 codes). The schemes designed for recorded transcript coding contained between 15 and 199 codes (median: 30 codes).

3.3.2 Use of new technologies

In line with latest developments, interviews may be recorded as a digital sound file. In this way the computer is not only used as device to go through a questionnaire (CATI or CAPI), but also enables 'Computer Audio Recorded Interviewing' (CARI), using the computer as a 'sophisticated tape recorder' (Biemer et al. 2000). Because no additional recording device such as a tape recorder is visible, recording is less obtrusive and respondents and interviewers are more likely to forget about the recording during the interview. With CARI the software instead of the interviewer controls recording, and arrangement of recording (e.g., to start concurrently with the interview or skip recording at specific sections) can be integrated with CATI/CAPI software (see Ongena, Dijkstra and Draisma 2004).

As Shepherd and Vincent (1991) argue, when coders compare question wording with interviewer's wording "they need to review a questionnaire source document that is identical to the document used by the interviewer" (Shepherd and Vincent 1991, p. 529). Therefore, if interviews are conducted by means of computer-assisted interviewing, ideally an electronic version of the questionnaire should be available, e.g., to account for complex skip patterns and automatically adapted question wordings. In Shepherd and Vincent's study, the coders used the CAI program itself, in order to view the questionnaire in exactly the same way as how interviewers had it available during the interview. In the Sequence Viewer program (Dijkstra 2002), several sections on the screen are available for coders with information on the exact question wording, the response alternatives and show cards used in the interview.

3.3.3 Availability of code descriptions

In order to warrant the reliability of results, it must be clear to what kind of behaviors a coder should apply certain codes. Interpretation of results will certainly be difficult if coders did not uniformly understand when to apply which code. Of course it is impossible to provide descriptions of all possible ambiguous situations. Therefore it is useful to document extraordinary situations by letting coders make notes for the ambiguities they came across in coding. The researcher can subsequently use these notes to adapt instructions for all coders.

Authors often give only an overview of the codes they used, and only indicate the code with two or three words ('adequate answer', 'inappropriate probe' etc.). Some authors (e.g., Cannell et al.1975, Prüfer and Rexroth 1985, Snijkers 2002) present their codes more clearly

because they give a short description (e.g., 'makes up in own words a probe (query) which is non-directive').

Brenner (1982) is one of the authors who present his codes the clearest, by not only describing them, but also giving fragments of Q-A sequences to illustrate the codes. Dijkstra (1999) uses the same strategy with clear examples, which are essential to explain his multivariate coding scheme, meaning that each utterance is coded on several descriptive variables (see section 3.3.4).

3.3.4 Organization of the coding scheme

In case of a large number of codes, it is important that the coder is able to manage this number of codes, to quickly choose the right code. This management is obviously improved when codes are well organized, for example by means of grouping them into similar categories of behavior. These categories may also be a means to reduce the number of codes, when for some analyses the different codes within a category are treated as one category. Cannell et al. (1975) for example grouped their codes into limited sets of interviewer activities, such as 'posing questions', 'probing and clarifying', and 'other behavior'. These sets were each arranged in two groups of correct and incorrect behaviors. The codes consist of two digits, with the first digit indicating the code category (e.g., 'correct question reading') and the second a further specification (e.g., 'reading the question exactly as worded'). It is therefore possible to use a reduced version of the coding scheme, using only the first digit.

In Dijkstra's (1999) *multivariate* coding scheme the behaviors of the interviewer and respondent are coded on a number of different coding variables. The coder, accordingly, needs to make several decisions (i.e., for each variable) when coding one utterance. Instead of making one decision concerning the choice between up to hundred different codes, as in the schemes of Blair (1978) and Prüfer and Rexroth (1985), the coder makes the same decision in multiple small steps. Using this procedure, the coders need to memorize only a relatively small number of codes, whereas the combination of the code variables may result in a very large amount of different codes. A multivariate scheme may be more reliable than an univariate one, because when coders choose wrong code values on one variable, the other variables may be correctly coded (Dijkstra 2002). Loosveldt (1985) used a similar strategy, and also Mathiowetz and Cannell's (1980) and Blair's (1978) coding schemes can be considered as multivariate.

3.3.5 The coders

The validity and reliability of the results obtained with the coding scheme depends on the persons who did the coding. As experimental research in social psychology has shown, observers may draw on specific theories when assigning meaning to behavior. For example, observers are more likely to draw on what they know about the actor's character in explaining behavior than when they explain their own behavior (for a review of experimental

studies see Watson 1982). Coders need to be trained especially in case of complex coding schemes.

Coders may be biased by researcher's expectations and make inferences based upon these expectations. Bakeman and Gottman (1997) state that it is important not to inform coders about hypotheses of a behavior coding study. In addition, they point out that not only inter-coder reliability is important, but also intra-coder reliability. Especially in case of complex coding schemes and when the coding process takes a long time, the coding may loose consistency. Moreover, it can hardly be avoided that coders develop their own expectancies during coding. A useful check is to compare codes assigned during the first half of the coding work with the second half.

Researchers

Some researchers (Brenner 1982; Loosveldt 1985; Van der Zouwen and Smit 2004) did the coding themselves, almost turning behavior coding into some kind of expert review. Apparently they only trust themselves in grasping the subtleties of such coding schemes. As Brenner (1982) states: "it proved impossible to find people who were willing, against payment, to code the tapes to a sufficiently high standard" (Brenner 1982, p. 143).

A disadvantage of this strategy is that not only coding may be biased by the researchers' hypotheses about the outcomes, but also that the coding scheme may be less appropriate to be used reliably by other researchers. Therefore, reliability scores of studies with researchers doing the coding themselves should be interpreted with care.

Field staff

A second possibility is to use field staff: either experienced interviewers who did not participate in the survey being coded, or supervisors, "control staff", "researchers" or "methodologists" as coders. An advantage of using this group is that these persons are familiar (or ought to be familiar) with interviewing conventions.

In the studies of Burgess and Patton (1993) and Snijkers (2002), the interviewers participating in the survey did the coding (of respondent behavior) themselves during the interview (using 5 and 7 different codes respectively). According to Burgess and Patton, coding could be applied easily, as 'proven' by perceptible delays in the interviews of 'only' 2-3 seconds for each code to be entered which "added perhaps 10 seconds on average to the length of the interviews, which averaged over 30 minutes" (Burgess and Patton, 1993, p. 396). In Burgess and Patton's (1993) study less than 3% of the Q-A sequences received a code. However, it is very unlikely that the target behaviors (i.e., respondent asks for repetition or clarification, interrupts interviewer, asks the remaining time left for the interview, or seemed uncomfortable) occurred in only 3% of the Q-A sequences, and therefore this clearly illustrates that an interviewer is not capable of capturing all occurrences of behaviors that need to be coded. Moreover, the fact that interviewers are coding the respondent's behavior may itself influence the interaction, as suggested by a side effect that

Snijkers (2002) found: it appeared to make interviewers more alert to problems with questions.

Trained coders

A third group of coders are specially trained coders, who do not necessarily have interviewing experience. Unlike using interviewers as coders, these coders should also be trained with respect to interviewing conventions.

Coders may be provided with verbal descriptions of the coding scheme and its application, followed by practical sessions with feedback from the researcher (Sykes and Collins 1992), or a manual with exercises (Dijkstra et al. 1985). The length of training may vary from one to two hours individual training (Blair 1978) to 45 hours (Oksenberg, Cannell and Blixt, 1996). Training of coders may also take place with a simultaneous further development of the coding scheme (Belli et al. 2004).

3.4 Reliability of the coding scheme

In 23 studies reliability scores are presented. Unfortunately, researchers do not use the same methods of determining reliability. Moreover, they do not all present their methods clearly, and therefore we can often only guess how reliability scores were produced and computed.

Reliability checks should be done with samples of multiple interviewers and respondents. It is better to double code random parts of multiple interviews than to double code one or a few complete interviews, because both interviewer and respondent styles may differ greatly, and more differences between interviews will be found than within one interview (Cannell et al. 1975).

Generally, the best way to test reliability is to test it at the same level as the level that was used for assigning codes. The more general the level, the less informative reliability scores are. For example, when codes are applied at the Q-A sequence level, we only know if coders agree that a certain behavior occurred in a Q-A sequence, but do not know whether or not coders based this decision on the same utterance. It is perfectly possible that multiple instances of the same behavior take place within the same Q-A sequences. Therefore, reliability scores at the Q-A sequence level, are generally overestimated.

Agreement scores at the utterance level can be divided into two different types: agreement upon what should be considered a separate utterance and agreement upon the individual codes s(Smit 1995). However, in most behavior coding studies reliability of these two types of agreement is rarely established.

Researchers are not uniform in statistics used for reliability testing (i.e., Kappa statistics, Pearson correlations or simple percentages). Percentages of agreement are computed by dividing the number of units with the same code by the total number of units coded. When the coding scheme contains only few different codes, the probability of chance agreement is very high. In the Kappa statistic the probability of chance agreement is incorporated.

In a number of cases the authors give detailed reliability information, e.g., separate reliability scores for interviewer and respondent behaviors, or even for each separate code category, which in some cases is quite low (c.f. Blair 1978, Oksenberg et al. 1991, Belli et al. 2004, Edwards et al. 2004). A low reliability score may not only be the result of ambiguity between two or more different code categories, but also of the absence of adequate code descriptions, inadequately skilled coders, or an inappropriate coding procedure.

The negative relation between code complexity and accuracy is often demonstrated (e.g., see Dorsey, Rosemery and Hayes 1986). Intuitively it makes sense that accuracy and inter-observer agreement are higher when the coding task is simpler. However, the correlation between the number of codes included (as a measure of coding scheme complexity) and the overall reliability score of Kappa values appeared to be positive but non-significant ($r = .166$, $p > 0.05$), for the 16 coding schemes for which kappa measures were available. Neither were differences in reliability scores related to the strategy (full, selective, sequential), the procedure (transcript coding, live or recorded coding) or the kind of coders used.

3.5 Focus of the coding scheme

Bakeman and Gottman (1997) state that creating a coding scheme is theoretically based, because the coding scheme represents a hypothesis. The scheme contains behaviors and distinctions that a researcher considers as being important. Therefore, they argue that researchers can only rarely use coding schemes of others. A different research question indicates a different coding scheme, and this would imply that comparing coding schemes developed for different research questions is not useful.

However, this might be less true for coding schemes designed to describe the behavior in standardized survey interviews. As Table 3-1 already indicated, quite some overlap can be found in the codes included in the 48 coding schemes. Virtually all behavior coding schemes describe the basic behaviors in an interview and at least have the implicit or explicit goal of finding departures from the paradigmatic sequence in common. The behaviors are usually evaluated in terms of ‘adequate’, ‘neutral’ or ‘inadequate’. However, depending on specific research questions, coding schemes often differ considerably from each other with respect to finer discriminations. For example, a scheme may be developed to evaluate behavior in a specific type of interview (such as the Event History Calendar, see Belli et al. 2004).

Based upon the elements of the data collection process that in one way or another may affect the response obtained, we define four different foci of a coding scheme; interviewers, respondents, questions and the interaction. These elements are partly derived from Cannell and Oksenberg’s (1988) distinction of goals of behavior coding. Studies can serve a meta-methodological goal (i.e., comparing different coding schemes or comparing behavior coding with other evaluation or pretest methods). However, the coding schemes in those meta-methodological studies can themselves always be classified according to the original focus, i.e., the element(s) they serve to pretest or evaluate. Schemes can also have multiple foci (e.g., Belli et al. 2004; Cannell et al. 1968).

In order to compare the different studies with respect to the aspects as discussed in the previous sections, and relate these aspects to the focus of the study, we will use a number of different categories that summarize the main characteristics of the coding scheme (see Table 3-6). We distinguished between three different aspects: the coding strategy, practical considerations in the coding procedure and the reliability of the scheme. Combining the two aspects of the coding strategy yields four different strategies: (a) selective coding at the Q-A sequence level (with no sequential information), which is often referred to as 'conventional behavior coding', (b) selective coding at the exchange level, (c) selective coding at the utterance level, and (d) full coding with sequential information, which is often referred to as 'interaction coding'. The strategies (b) and (c) yield sequence information only at the exchange level, and therefore these two categories are integrated as one category. Additional aspects of a coding scheme are the number of actors involved (i.e., interviewer, respondent and possible third parties), the number of codes included, the mode of administration (face-to-face or telephone).

Table 3-6 Overview of aspects of comparison of behavior coding schemes

Aspect	*Abr.*	*Specification*
Coding strategy	SN	Selective coding at the Q-A sequence level, no sequence information
	SE	Selective coding with sequence information at exchange levels
	FS	Full coding with sequence information preserved
Coding procedure	L	Live coding
	Lr	Live coding, recording on cassette as backup
	Rc	Recorded tape coding
	Rt	Recorded transcript coding
	Rc/t	Recorded tape coding with transcripts as backup
Reliability procedure	K	Kappa
	KD	Kappa with unit of analysis *deviating* from level of coding
	P	Percentage
	PD	Percentage with unit of analysis *deviating* from level of coding
	C	Pearson correlation

3.5.1 The interviewer as a focus: interviewer monitoring studies

As Cannell and Oksenberg (1988) point out, results of interviewer monitoring studies can be used in terms of supervision ('enforcing rule following behavior') and evaluation (assessing quality of particular studies, assessing overall staff performance, evaluating training methods, or exploring ways to improve training).

It appears that especially many of the early behavior coding schemes are designed for the goal of monitoring interviewer performance (i.e., 14 of the 48 schemes compared). Table 3-7 shows that most coding schemes that were designed for interviewer monitoring use a selective coding scheme that does not preserve sequential information, and none of them uses a full coding scheme. Furthermore, many interview monitoring schemes include only interviewer behavior codes, such as Cannell et al.'s (1975) scheme that served as a basis for

many coding schemes (also for coding schemes with different foci, i.e.: Morton-Williams 1979; Prüfer and Rexroth 1985; Sykes and Collins 1992). Their scheme included all concepts and principles that were considered to be important targets in interviewer training. From this viewpoint the interviewer and respondent were considered as individual actors, which individually could produce errors, unrelated to each other.

Table 3-7 Coding schemes with interviewer behavior as focus

Scheme	Coding	Actors	Number of different codes		Procedure	Mode	Reliability procedure	Overall reliability
			I	R				
Fowler and Mangione (1990)	SN	I	11	-	-	-	-	-
Couper, Holland & Groves (1992)	SN	I	16	-	L	T	-	-
Mathiowetz & Cannell (1980)	SN	I	20	-	Lr	T	P	.88
Cannell, Fowler & Marquis (1968)	SN	IR	5	7	L	F	-	-
Bradburn & Sudman (1979)	SN	IR	4	2	Rc	F	K	.52-.72
Blair (1978)	SN	IR	39	11	Rc	F	K	.74
Blair (1980)	SN	I	4	-	Rc	F	-	-
Shepherd & Vincent (1991, 'compliance')	SN	I	16	-	Rc	T	-	-
Oksenberg, Cannell & Blixt (1996)	SN	IR	14	7	Rc		K	.11-.90
Stanley (1996)	SN	IR	5	6	Rc	F	-	-
Brick, Tubbs et al. (1997b)	SN	IR	5	6	Rc	T	P	.48-.68
Carton (1999)	SN	IR	41	12	Rc	F	-	-
Cannell, Lawson and Hausser (1975)	SE	I	30	-	Rc	T	K	.80-.92
Prüfer & Rexroth (study 1, 1985)	SE	I	35	-	Rc	F	-	-
Belli et. al (2004)	SE	IR	25	17	Rt	F	C	.42-1.0

Coding strategy: SN = selective, no sequential info, SE = selective, sequential at exchange levels, *Actors*: I = interviewer, R = respondent, *Coding procedure*: L = live coding, Lr = live coding backup tape, Rc = Recorded tape coding, Rt= recorded transcript coding, *Mode*: F= face-to-face, T = telephone, *Reliability procedure*: K= kappa, P = percentage, C = correlation (one overall reliability score or the minimum and maximum of all scores)

Codes included

Typically, interviewer monitoring schemes include the quality of question reading (distinguishing exact reading from reading with minor and/or major changes) and adherence to skip patterns. This *unconditional* scripted behavior mainly occurs before the respondent has spoken, and therefore interviewers usually have direct control over this behavior. Belli and Lepkowski (1996) conclude that "respondent behavior is more diagnostic of response accuracy than anything over which the interviewer has direct control" (Belli and Lepkowski, 1996, p.73). Therefore, it is very useful to also include codes that evaluate the interviewer's reaction to respondent behavior, i.e., *conditional* (un)scripted behavior. Furthermore, more than half of these coding schemes also include respondent behavior codes, which may be very relevant to evaluate interviewer behavior, e.g., to determine whether interviewers appropriately reacted to certain respondent behaviors.

Alternative methods

Alternative assessments of interviewer's work (i.e., reviews of completed questionnaires and response distributions), although inexpensive and easily conducted, appear to reveal only a small part of inadequate interviewer performance (see Wilcox 1963, cited by Cannell and Oksenberg 1988). Such methods leave errors in the most important interviewer tasks (reading questions and probing) undetected. Direct observation by a supervisor is usually subjective and unsystematic, and therefore, as Cannell and Oksenberg state, "standardized coding of interviewer behavior provides an objective method for evaluating interviewer performance" (Cannel and Oksenberg, 1988, p. 475).

3.5.2 The questions as a focus: evaluating questions

Another focus of a behavior coding scheme is to identify questions that cause problems for the interviewer or respondent, in order to pretest, evaluate or explore the effects of question wording. This focus has become more important since the first scheme of Morton-Williams (1979), and is the most frequently used focus of behavior coding schemes (i.e., 21 of the 48 schemes, see Table 3-8).

The rules that are the basis for these schemes and the codes that result from them concern (problematic) interviewer as well as respondent behavior. For example, Morton-Williams (1979) gives nine criteria for adequate question performance. The categories of interviewer and respondent behavior she subsequently describes refer to the criteria on which a question might have failed.

Table 3-8 Coding schemes with the questions as focus

Scheme	Coding	Actors	Number of different codes		Procedure	Mode	Reliability procedure	Overall reliability
			I	R				
Burgess & Patton (1993)	SN	R	-	5	L	T	-	-
Snijkers (2002)	SN	R	-	7	L	TF	-	-
Presser & Blair (1994)	SN	IR	2	3	L	T	-	-
Oksenberg, Cannell & Kalton, (1991)	SN	IR	3	7	-	-	K	.60-.80
Edwards et al (2002)	SN	R	-	2	L/Rc	T	K	.38
Blixt & Dykema (1995)	SN	R	-	5	Rc	F	K	.65
Sykes & Morton-Williams (1987)	SN	IR	1	5	Rc	F	-	-
Sykes & Morton-Williams (1987)	SN	IR	2	8	Rc	F	-	-
Gustavson, Herrman & Puskar (1991)	SN	IR	9	6	Rc	F	K	.55-.82
DeMaio et al. (1993)	SN	IR	6	6	Rc	TF	-	-
Cahalan et al. (1994)	SN	IR	15	8	Rc	T	-	-
Dykema et al. (1997)	SN	IR	4	6	Rc	F	-	-
Hess, Singer & Bushery (1997)	SN	IR	5	8	Rc	T	K	.55-.85
Lepkowski et al. (1998)	SN	IR	6	13	Rc	F	K	.18-.77
Bates & Good (1996)	SN	IR	4	5	Rt	F	P	.83
Van der Zouwen & Smit (2004)	SN	IR	8	7	Rt	F	KD	.76
Edwards et al (2004)	SN	IR	9	9	Rc	T	K	0.0-1.0
Esposito et al. (1992)	SE	IR	6	7	L	T	PD	.87
Schaeffer & Dykema (2004)	SE	IR	15	14	Rc	T	KD	.75+
Morton-Williams (1979)	SE	IR	14	17	Rc	F	-	-
Prüfer & Rexroth 1985 (study 2)	FS	IR	59	28	Rc	F	-	-

Coding strategy: SN = selective, no sequential info, SE = selective, sequential at exchange levels, FS = full coding, *Actors*: I = interviewer, R = respondent, *Coding procedure*: L = live coding, Rc = Recorded tape coding, Rt= recorded transcript coding, *Mode*: F= face-to-face, T = telephone, *Reliability procedure*: K= kappa, Kd=kappa deviated level, P = percentage, Pd=Percentage deviated level, C = correlation (one overall reliability score or the minimum and maximum of all scores)

Codes included

Typically, coding schemes to evaluate questions include interviewer's question reading codes and respondent codes that are assumed to indicate problems in question understanding such as interruptions, requests for clarification, qualified and inadequate answers. However, these behavioral categories occur quite infrequently. As Schaeffer and Maynard (2002) also suggest, a number of other behavioral categories not typically included in behavior coding schemes (e.g., hesitations, reports, and feedback codes) may be much more effective in signaling problems with question wording, especially as compared to explicit requests for clarification, which respondents may avoid to use, as a result of standardized interviewing practice.

Sources of problematic behaviors can often be found by means of comparison of the percentage of problematic behaviors and the specific question wording (Oksenberg et al.,

1991). Close inspection of question wording may reveal why interviewers frequently change it, or why respondents frequently interrupt it or give qualified answers. Furthermore, additional cues may be derived from answer distributions, information from coders and interviewers, and the transcripts, if available. Fowler's (1992) study illustrated the usefulness of behavior coding as a diagnostic tool that is also helpful to suggest revisions of question wording that improve the validity of data. Questions, which were identified as problematic by means of behavior coding, were redesigned. The alternative question wording not only yielded fewer instances of problematic behaviors, but also response distributions that were expected to be more accurate.

Alternative methods

Behavior coding has often been judged as less effective in diagnosing problems with question wording than for instance cognitive interviewing (for a review, see Campanelli 1997). However, comparisons typically involve behavior coding in its usual implementation i.e., in simple 'selective' (see section 3.2.2) coding schemes using only common codes such as those listed in Table 3-1.

In articles that compare behavior coding with other methods for their sensitivity of detecting problematic questions it is often recommended to use combinations of techniques, each yielding unique contributions (DeMaio et al. 1993; Hughes 2004; Presser and Blair 1994; Willis, Schechter and Whitaker 1999). Furthermore, it is rather difficult or even useless to compare pretesting methods. Cognitive interviews have their own disadvantages, e.g., they can influence the question-answering process they seek to explore, because (especially concurrent) think-aloud instructions disturb the actual question-answer process. Moreover, respondents are not always able to spontaneously express their cognitive processes, especially when retrospective think aloud is applied (see Sudman, Bradburn and Schwarz 1996). With cognitive interviews, the chances of finding non-existing problems are larger, but the chances of not finding existing problems are smaller than with behavior coding. However, behavior coding often is the only method that evaluates the interviewer objectively. Furthermore, behavior coding often is the only method that is quantitative and easy to replicate. Therefore, cognitive interviewing is ideally implemented in an operationalization phase, whereas behavior coding is ideally implemented in a pilot study of pretesting questions (Willis 2005).

3.5.3 The respondent as a focus

Monitoring respondent performance as a focus may seem odd at first sight, because a supervisor can hardly correct respondents. However, a researcher can monitor the behavior of respondents in survey interviews in order to identify and describe difficult to interview respondents, or aim to find questions that are difficult for particular respondents. Four schemes that were (partly) designed to monitor respondent performance are summarized in Table 3-9.

Table 3-9 Coding schemes with respondent behavior as a focus

Scheme	Coding	Actors	Number of different codes		Procedure	Mode	Reliability procedure	Overall reliability
			I	R				
Cannell et. al. (1968)	SN	IR	5	7	L	F	-	-
Loosveldt (1997)	SN	R	-	6	Rc	F	-	-
Gallagher, Fowler and Roman (2004)	SE	IR	15	9	Rc	T	-	-
Belli et. al (2004)	SE	IR	25	17	Rt	T	C	.42-1.0

Coding strategy: SN = selective, no sequential info, SE = selective, sequential at exchange levels, *Actors*: I = interviewer, R = respondent, *Coding procedure*: L = live coding, Rc = Recorded tape coding, Rt= recorded transcript coding, *Mode*: F= face-to-face, T = telephone, *Reliability procedure*: C = correlation (the minimum and maximum of all scores)

Loosveldt (1997) used six respondent behavior categories as objective indicators of the respondent's cognitive and communicative skills. Gallagher et al. (2004) tested the effects of training aged respondents in their role, which appeared to be effective to reduce the number of interruptions, but not with respect to reducing interview length, response rates, or refusal rates.

Alternative methods

Alternative assessments of response quality (i.e., item non response and biases in response distributions), will reveal (like similar measurements to assess interviewer performance) only a small part of inadequate respondent behavior. Methods like interviewer debriefing or direct observations are also likely to be incomplete and subjective.

3.5.4 The interaction as a focus

Another goal of behavior coding studies can be to examine what the effects of specific behaviors will be on subsequent behaviors, or the interactional causes of specific behaviors. Hill and Lepkowski (1996) use the term 'behavioral contagion' to indicate their goal to study how one instance of deviating behavior can lead to another instance.

Although all schemes that include evaluations of both interviewer and respondent behaviors may provide knowledge about the interaction, what is different about interaction schemes, is the sequential information that is analyzed (i.e., in which order the behaviors occurred). As is shown in Table 3-10, all schemes include (some) sequential information. Furthermore, the table shows that all schemes include 20 or more codes, except for Hill and Lepkowski's (1996) scheme. The studies often have an explorative character (c.f. Lepkowski et al. 2000, Sykes and Collins, 1992)

Table 3-10 Coding schemes with the interaction as a focus

Scheme	Coding	Actors	Number of different codes		Procedure	Mode	Reliability procedure	Overall reliability
			I	R				
Hill & Lepkowski (1996)	SE	IR	2	4	Lr	F	-	-
Shepherd & Vincent (1991,' interaction')	SE	IR	21	18	Rc	T	-	-
Lepkowski et. al. (2000)	SE	IR	14	9	Rc	F		-
Sykes & Collins (1992)	SE	IR	35	19	Rc	F	PD	.88
Marquis & Cannell (1969)	FS	IR	27	20	Rc	F	-	-
Loosveldt (1985)	FS	IR	95	79	Rc	F	-	-
Brenner (1982)	FS	IR	18	6	Rc/t	F	-	-
Smit (1995)	FS	IR	10	10	Rt	F	K	.72
Dijkstra et. al. (1985) & Dijkstra (1983)	FS	IRP	24	15	Rt	F	K	.80
Dijkstra (1999; 2002)	FS	IRP	±139	±60	Rt	-	K	.78

Coding strategy: SE = selective, sequential at exchange levels, FS = full coding, *Actors*: I = interviewer, R = respondent, P = third person, *Coding procedure*: Lr = live coding backup tape, Rc = Recorded tape coding, Rt= recorded transcript coding, Rc/t = Recorded tape coding, backup transcript, *Mode*: F= face-to-face, T = telephone, *Reliability procedure*: K= kappa, Pd=Percentage deviated level (one overall reliability score)

Marquis and Cannell (1969) conducted the first interactional study. As early as 1968, Cannell et al. reflected on a 'reciprocal cue searching process' to be present in interviews. Their data led them to speculate about the existence of a process during the interview in which both interviewer and respondent are searching for cues about appropriate kinds of behavior (Cannell et al. 1968). Because the data in this study did not allow interactional analyses to prove these speculations to be right, in 1969 Marquis and Cannell used a revised coding scheme and coding procedure. They performed analyses on for instance the effects of directive and neutral probes on respondents giving adequate answers, or the probability that interviewer feedback follows specific categories of respondent behavior. Brenner (1982) recognized the importance of studying interactional processes (which he called 'action-by-action' analysis) and was among the first who performed such analyses.

Codes included

An important difference with respect to the codes included in schemes for interaction analysis as compared to other schemes, is that usually non-problematic behaviors are also coded (i.e., reports, elaborations, perceptions, comments, etc.) in order to more fully describe the interaction. However, this requires a complex coding scheme, and not all non-problematic behaviors may be relevant. Using a summary code ('other behavior') can compensate for this problem. Although such a code will reduce the information available, it is possible to distinguish sequences with these summarizing codes from paradigmatic sequences.

Therefore, it is always possible in a later stage to recode the summary codes into finer distinctions, if necessary.

Alternative methods

Behavior coding suffers from the bias that it should be determined in advance what behaviors are relevant. Even full coding schemes do not always make fine discriminations, and may neglect distinctions that might be relevant afterwards. Therefore qualitative methods, such as conversation analytic studies may be useful. In that case transcripts are required, often using a detailed method of transcription according to conversation analysis conventions, as developed by Gail Jefferson (1983) .

However, by using a full coding scheme with sequential information, the original question wordings, and the entered responses, a lot of information is available to researchers. This may not only fulfill the requirements of availability of detailed data, it also enables a quantification of such data. Thus, behavior coding enables a researcher to determine whether odd interactions are unusual incidents or evidence that data obtained by standardized interviews is untrustworthy. In this way, behavior coding may be helpful in resolving discussions between practitioners and critics of standardized interviewing (Maynard and Schaeffer 2002). Quantification is precisely what is lacking in qualitative data analysis, and therefore often used as a critique towards qualitative studies, as Houtkoop-Steenstra (2000) also notes.

3.6 Conclusion

The verbal behavior in a standardized interview yields a wealth of information that can be used for various goals. Because the behavior takes place in structured sequences of questions and answers, most coding schemes have many common elements.

In Table 3-11, the coding strategies and schemes that are recommended for different situations are listed. The choice of the coding strategy depends to an important extent on the focus and goal of the scheme. In case of pretesting and monitoring (relevant parts of) the data collection, it is important that quick results are available, in order to enable efficient processing of adaptations. This behavior coding takes place *prior to* or *during* actual data collection. Therefore schemes appropriate in this initial phase are limited to selective schemes (with less than 15 codes) that aim for frequency analysis.

Performing behavior coding for evaluation or exploration (of relevant parts) of the data collection process can take place *after* actual data collection. In case of evaluation quick results are not important, but detailed explanations of causes of problematic behaviors may not be relevant, and therefore selective coding schemes with slightly higher number of codes (i.e., around 20) may be appropriate.

In case of exploratory analyses of the interaction, detailed information is required, and full coding schemes with sequential information seem most appropriate. However, for

practical application of such schemes, software like the Sequence Viewer program[6] is necessary (see Dijkstra 2002).

Table 3-11 Coding schemes for specific phase, goal and type of analysis

Focus	Type of study	Strategy	Type of analysis	Procedure	Examples of schemes
Interviewers	Monitoring	Selective	Frequency	Live	Couper et al. (1992)
	Monitoring	Selective	Frequency	Tape	Brick et al. (1997), Stanley (1996)
	Evaluation	Selective	Frequency	Tape	Oksenberg et al. (1996)
	Experiment	Selective	Frequency	Tape	Cannell et al (1975)
Questions	Pretest	Selective	Frequency	Live	Presser and Blair (1994)
	Pretest	Selective	Frequency	Tape	Oksenberg et al (1991), DeMaio et al (1993)
	Evaluation	Selective	Sequence (exchange)	Tape	Lepkowski et al. (2000), Morton-Wiliams (1979)
	Exploration	Selective	Sequence (exchange)	Tape	Schaeffer and Dykema (2004)
	Experiment	Full	Sequence (utterances)	Transcript	Dijkstra (1999)
Respondents	Evaluation	Selective	Frequency	Tape	Gallagher (2004)
Interaction	Exploration	Full	Sequence (exchange)	Tape	Sykes and Collins (1992)
	Exploration	Full	Sequence (utterances)	Transcript	Dijkstra (1999)

The goal of interaction analysis performed in this thesis is to add to the theoretical knowledge of processes in the interview, and to provide suggestions for procedures to improve the quality of data collected by means of survey interviews. We aim to detect systematic problems that occur in the interaction between interviewers and respondents, and also to identify the causes of these problems. Thus, we need a coding scheme that is detailed enough to detect those problems, and their interactional causes, but that is feasible enough to enable coding of high reliability and validity and a not too complex to perform clear methods of analysis. In the next chapter a coding scheme will be described that fulfills these criteria.

[6] This program can be used to store, transcribe, code and analyze interviewer-respondent interactions. The program will be described with more detail in section 4.2.1, chapter 4

4 Description of the coding scheme

4.1 Introduction

Our research questions are concerned with problematic deviations in Q-A sequences and causes of those deviations. To answer these questions we need a detailed coding scheme. In the previous chapter we gave an overview of methods for behavior coding, and compared several coding schemes. We concluded that a full coding scheme with preservation of sequential information is the best method to perform interaction analysis in a quantitative way. In this chapter, we describe the coding scheme that met these and other criteria, and is used in this thesis. We explain the choices we made with respect to coding options included in the scheme. We first give a general description of the coding scheme. Secondly, we will discuss the coding options we use in our scheme in more detail, and relate them to the categories used in other schemes, if appropriate.

4.2 The coding scheme used in this thesis

Our research questions are exploratory; and they do not concern specific types of behavior. Empirical studies that attempt to describe the interaction in a systematic and detailed way hardly exist. Thus, we do not know in advance what types of behaviors may be most relevant to describe the course of the interaction. However, theoretical perspectives and other empirical studies may indicate what behaviors are relevant for the quality of the responses obtained. To recapitulate our discussion of conversational and cognitive perspectives on the interaction in chapter 2, relevant behaviors may be:

- Comments of respondents (which indicate 'task involvement', see section 2.2.4)
- Behavior of third parties (interfering the interaction between interviewer and respondent, see section 2.2.5)
- Inference (e.g., when interviewers have knowledge of a survey topic they may be less likely to probe after inadequate answers, see section 2.2.6)
- Interviewer's rewording of questions (e.g., adaptations according to politeness, appropriate length and structure, etc., see sections 2.2.8 and 2.4.2)
- Interruption (caused by conventions in turn taking see section 2.2.9)
- Acknowledgments (repeats, neutral perceptions or assessments, see section 2.2.10)
- Initial and later responses (i.e., initial responses may be produced with hesitance, see section 2.2.11)
- Suggestive probes (respondents are likely to accept suggestions, see section 2.2.11)
- Skipping questions (respondents may provide information that is not verified, see section 2.2.12)
- Explicit and implicit request for clarification (which can have consequences for adherence to standardization rules by interviewers, see section 2.2.13)
- Interviewer's clarification of questions, repetition of question, 'Whatever it means to you replies' (which can have consequences for standardization of meaning, see section 2.2.13)

- Preciseness of formatted answer (i.e., mismatch answers, which require probing, see sections 2.2.14 and 2.3.3)
- Elaborations (which distract interviewer and respondent from their task, see section 2.2.15)
- Verbal expressions of uncertainty and enumerations (which may cause interviewers to infer answers, see section 2.3.2).

Most of these behaviors are covered in the 48 coding schemes that were discussed in chapter 3. However, also behaviors that are not expected to influence the quality of the response obtained may be relevant to obtain a complete description of the interview. We also have only general ideas about what behaviors occur most frequently, and in which order, and who is responsible for the first problematic deviation from the paradigmatic Q-A sequence. Thus, we need a rather detailed coding scheme. When we would use a coding scheme with only a selection of categories, or too global categories, it is possible that important behaviors are overlooked. Furthermore, a detailed coding scheme enables better possibilities to find theoretical explanations for the course of the interaction.

However, a detailed coding scheme may require a large number of codes. Such a scheme is very difficult to manage for coders. Moreover, a large number of codes may create a complex dataset that is difficult to analyze. Preferably, initial exploratory analyses are done in rough categories, to first identify the most common problematic behaviors. In subsequent analyses the subtle distinctions within those rough categories and relations with other behaviors can be distinguished. Therefore it may be useful to organize the subtle distinctions in comprehensive categories. The multivariate coding scheme of Dijkstra (1999) is precisely organized in such a way. The principle of a multivariate coding scheme is that utterances are coded for multiple coding variables. Each coding variable describes a particular aspect of the utterance. The combination of values yields a code string that constitutes a meaningful description of the utterance. For example, a code string can be ‘RRd’, which means that the respondent (R) requests (R) repetition (‘duplication’) of the question (d), or ‘RRm’, which means that the respondent (R) requests (R) clarification of the question’s meaning (m).

Suppose we want to investigate interactional causes of respondents’ requests for clarification or repetition of the question. However, we want the first analysis to be simple, and assume that there is no difference in the causes of these two types of requests (for clarification and repetition). To simplify the analysis, we can temporarily ignore whether the request refers to repetition (‘d’) or to clarification (‘m’) of the question. Thus, we ignore the coding variable describing the character of the request, restricting the code to ‘RR’ (‘respondent submits a request’). Suppose we find that respondents submit requests more frequently when the interviewer read the question as worded than when the interviewer changed a few words in question reading. When we want to further analyze causes of requests, we may find that the distinction between requests for clarification and for repetition becomes relevant. As a fictitious example, changing a few words in the question may appear to decrease chance of requests for repetition, but does not affect the occurrence of a request

for clarification. With a multivariate coding scheme it is simple to switch from analyses with only a few categories by excluding one or more coding variables, to detailed analyses with a large number of categories, when all values of all variables are included.

The coding scheme also has a cumulative hierarchical character, which means that once earlier coding variables (i.e., the most general ones) are coded, the possible values of subsequent variables are restricted to a subset of appropriate codes. For example, with a general coding variable, 'requests' are distinguished from 'answers'. When an utterance has been coded as 'request', the coders choice for the next, more specific variable has been restricted to just 'clarification' or 'duplication'. When an utterance has been coded as an 'answer' however, the coder has the choice from a different subset of codes to specify the particular type of answer. In this way, the coding task is simplified, and only code strings of logical combinations are obtained.

Both the division of the coding task into small subtasks, and the hierarchical character of the scheme reduce the chances of errors made in coding. Furthermore, it is possible that only part of the coding variables is coded inadequately. For example, the respondent's utterance may have been correctly recognized as a request, but may be falsely recognized as a request for clarification instead of duplication. Hence, for analyses excluding the specification, the coding is still reliable with respect to distinguishing requests from other types of utterances on the same level (e.g., 'questions' or 'answers'). Thus, the multivariate character of Dijkstra's scheme indicates that the scheme fulfills our criterion of practical feasibility and reliability.

The coding variables and categories of the scheme also indicate that the scheme fulfills our criterion of completeness. We compared the coding options included in the 48 coding schemes that were discussed in chapter 3. Although the 48 coding schemes show much overlap with respect to the code categories, we nevertheless found 134 different categories for interviewer behavior, 78 different categories for respondent behavior, and 14 different categories for behavior of third parties. It appeared that virtually all these different categories could be covered by particular combinations of codes on the different coding variables from Dijkstra's scheme. In addition this scheme was especially designed to capture details in the interaction, thus also including codes for behaviors that were less meaningful with regard to the goal of the majority of the 48 coding schemes and hence not incorporated in those coding schemes. These codes are typically referring to interactional behaviors, e.g., codes referring to perceptive behavior, and filled pauses like 'uhs'. We conclude that the categories of the scheme that we used fulfill our criterion of completeness.

In the next section we will first describe the practical usage of the multivariate scheme. In the sections thereafter we will describe the categories of the most important coding options as included in the 48 schemes, and we relate those to the coding options of Dijkstra's scheme.

4.2.1 Practical usage of the scheme in the Sequence Viewer program

The coding scheme is mainly developed to be used in a computer program, such as the Sequence Viewer program (Dijkstra, 2002). This program can be used to transcribe audio files, subsequently code these transcriptions, and analyze the codes.[7] For each Q-A sequence, a sequence record is created, and audio files can be linked to these records. In Figure 4-1 a screenshot of the program shows how most important information from a sequence record appears on the computer.

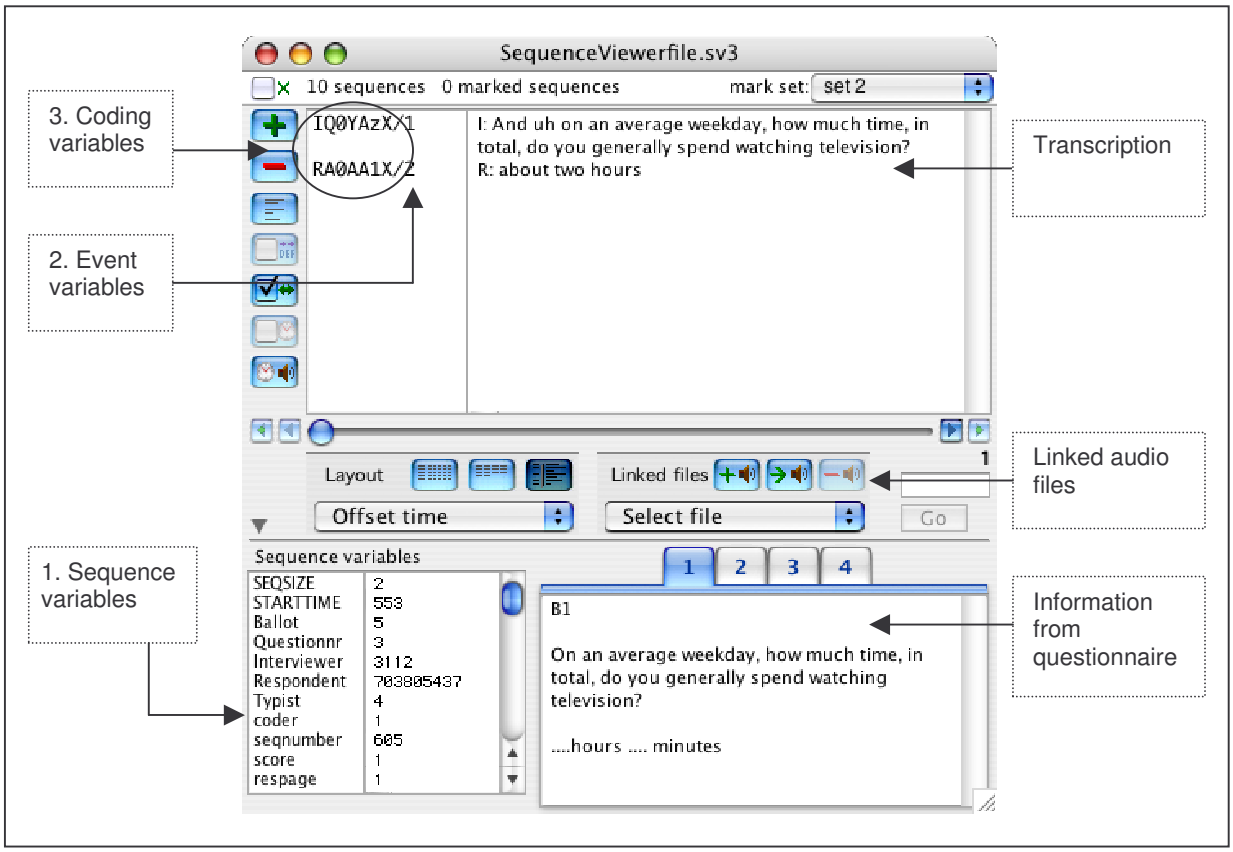

Figure 4-1 Screenshot of the Sequence Viewer program

Especially when interactions of standardized survey interviews are to be coded, a number of options in the program can be used to ease and speed up the transcription process. Both transcriptions and linked audio files (see Figure 4-1) can subsequently be used for coding the utterances. The program helps coders to choose among the code categories, by means of providing an overview of all possible code values within a coding variable. Semi-automatic coding (or even full automatic coding of paradigmatic Q-A sequences) can be applied by means of text recognition.

The program is organized around three types of variables (see Figure 4-1). The first type is called a 'sequence variable', and provides information about a Q-A sequence, questions, interviewers or respondents (e.g., the length of the sequence, the identification

[7] The Sequence Viewer program can be obtained free of charge from the website, http://home.fsw.vu.nl/w.dijkstra/SequenceViewer.html, but is as yet available for Macintosh computers only.

number of questions, interviewers and respondents, the score as entered by the interviewer, etc.). The second type is an 'event variable', and provides numerical information about the utterance (e.g., number of words in the utterance, duration, position in Q-A sequence, etc). The third type is a 'coding variable', which comprises the coding scheme.

4.2.2 Variables in the scheme

Five different coding variables are used in the coding scheme. Because both general and specific variables are used, it is possible to adapt the scheme to specific research questions (Dijkstra, 1999). In Appendix 4-1 an overview is given of these five coding variables and their complete set of values (and a sixth variable that is not relevant for this thesis).

Each utterance is coded on the five variables. Thus, codes are assigned to an utterance through a choice among values (abbreviated by means of one character, a number or a letter). A completely coded utterance consists of a string of (five) characters that each describe a particular aspect (i.e., a coding variable) of the utterance. These variables are labelled as ACTOR, EXCHANGE, DISTANCE, SPECIFICATION and ADEQUACY respectively. For example, a code string such as "RA0AA" means that the respondent (code 'R' on the variable ACTOR) gives an answer (code 'A' on EXCHANGE) on the question from the questionnaire (code '0' on DISTANCE) by means of a choice of one of the alternatives ('A' on SPECIFICATION), which is moreover an adequate answer ('A' on ADEQUACY). A summary of the most important categories for the five coding variables of the scheme is given in Table 4-1.

Table 4-1 Overview of most important categories of the coding scheme

ACTOR	EXCHANGE	DISTANCE	SPECIFICATION	ADEQUACY
I: interviewer R: P: third	Q: question	0,1,2	C: choice Q Y: yes-no Q A: alternatives I : introduction M: meaning of Q	A: adequate M: mismatch I: invalid S: suggestive
	A: answer	0,1,2	A: alternative O: open answer b: don't know r: refusal	A: adequate M: mismatch I: invalid T: qualified
	P: perception	0,1,2	E: echo n: notes other	A: adequate M: mismatch
	R: request	0,1,2	d: repetition m: meaning	x x
	C: comment	0,1,2	p: personal	x
	D: detour		t: task	x

ACTOR is a coding variable that indicates the producer of an utterance; i.e., the interviewer (I), the respondent (R), a third person (P) or any other actor that may be relevant in the interaction to be coded.

EXCHANGE is a coding variable that indicates the type of information that is

communicated in a general sense; like questions ('Q'), answers ('A'), or requests ('R'). Utterances may also be an indication that the other speaker was perceived (i.e., perception, code 'P'). The variable EXCHANGE may be viewed as a categorization according to the function of the utterance. It also comprises the main factor in the hierarchical organization of the scheme; it puts restrictions upon the possibilities of categories for subsequent coding variables, which are more specific variables, related to the evaluation of standardized survey interviews.

The variable DISTANCE comprises an evaluative component; the coder must decide to what extent the utterance is relevant to the question from the questionnaire. With this coding variable it is possible to indicate how far the interviewer or the respondent are digressing from the interview topic. When the interviewer or respondent are performing behavior that is directly related to the questions from the questionnaire (e.g., posing a question from the questionnaire, or answering such a question), the distance of their utterances is 0. When respondents elaborate and motivate (answers to) questions from the questionnaire (or the interviewer asks the respondent to do so, whereas such a question is unscripted) the distance is 1. Answers and questions with a distance 2 are elaborations, not related to the question anymore.

The variable SPECIFICATION may be used to further specify the category coded for EXCHANGE. For example, questions can be specified for the type of question (e.g., a choice question 'C' or an open question 'O'), answers can be coded as 'chooses an Alternative' (A), signifying that an utterance is perceived can be communicated in the form of an 'Echo' (repeats the other: 'E'), requests can be specified as request for repetition ('d'), or requests to clarify the meaning ('m'), and finally, both comments and detours can be specified as 'personal' (p).

The variable ADEQUACY evaluates the utterances with respect to standardization. Three general concepts are important here. The first concept is the *'validity'* of an utterance, which evaluates the correspondence with the meaning of questions. For example, a question is coded as 'invalid', if the interviewer changes the original meaning. An answer is 'invalid' if the respondent clearly misunderstood the question.

The second concept is *'mismatch'*, which evaluates the correspondence between the uttered questions or response alternatives and their scripted versions. For example, if the interviewer does not read the question as worded, but without changing this meaning (that would be an invalid question) the question is a 'mismatch' question. A mismatch answer is not exactly formatted as one of the scripted response alternatives.

The third concept is the *'suggestiveness'* of an utterance. This evaluation is only applied to utterances of interviewers (i.e., reading of questions or alternatives). Question reading is suggestive when one or a few response alternatives are offered, without previously received information from the respondent to warrant this selection.

In Example 4-1, a Q-A sequence is shown that is coded for all five coding variables

Example 4-1 Q-A sequence coded for all variables

	Verbal utterances	Explanation	ACTOR	EXCHANGE	DISTANCE	SPECIFICATION	ADEQUACY
1	I: Do you have a very good, good, reasonable or bad health?	I poses question adequately with all alternatives (1,2,3 and 4)	I	Q	0	C	A
2	R: Could you repeat that?	R requests repetition	R	R	0	d	x
3	I: Do you have a good, reasonable or bad health?	I poses question suggestively: only alternatives 2, 3 and 4.	I	Q	0	C	S
4	R: Not so bad	R gives mismatch answer	R	A	0	A	M
5	I: Would you say good, reasonable or bad?	I repeats alternatives adequately	I	Q	0	A	A
6	R: Well, reasonable	R gives direct answer, that is adequate	R	A	0	A	A
7	I: Reasonable	I adequately repeats direct answer	I	P	0	E	A
8	R: Compared to my wife yes	R elaborates his answer	R	A	1	O	x
9	I: So reasonable	I repeats direct answer	I	P	0	E	A

Having provided an overview of the coding scheme that is used in this thesis, in the next sections we will present in more detail which distinctions in codes were made, and how these relate to the codes that were identified across 48 different coding schemes. We will illustrate how coding options can be created by means of our multivariate coding scheme When particular coding variables are not relevant for illustration purposes, the code values will be indicated by means of '•'.

4.3 Interviewer behavior

No less than 134 different coding options that can be attributed to the interviewer were found in the coding schemes. A simple coding scheme may only evaluate 'reading the question', but of course the interviewer can perform numerous other actions. These actions can be evaluated on various aspects of adequacy. In Table 4-2 an overview is given of behavioral categories that can be distinguished for interviewer behavior, and examples of specific behaviors and evaluation criteria that are often applied in behavior coding schemes

Table 4-2 Interviewer behavioral categories and evaluation criteria

Type of interviewer behavior	*Examples*	*Evaluation criteria of adequacy*
Unconditional, scripted behavior	Question, Response alternatives, Introduction, Instruction	- Neutrality - Exact or Minor/Major change - Adherence to skip patterns
Conditional, scripted behavior	Repetition of Question or Response alternatives, Clarification, Probing, Feedback	- Neutrality - Exact/Minor/Major reading - Negligence - Redundancy
Conditional, unscripted behavior	Clarification, Probing, Feedback	- Neutrality - Negligence - Redundancy
Unscripted behavior	Irrelevant behavior	- Relevance
Non verbal behavior	Recording Answers	- Errors

4.3.1 Unconditional scripted interviewer behavior

Unconditional scripted behavior mainly occurs before the respondent has spoken, and may also be referred to as 'initial question reading'. It may comprise several specifications of interviewer behavior, such as reading of the question proper, introductions, response alternatives, etc.

All unconditional scripted behaviors receive the same codes for the first three coding variables of our coding scheme, i.e., 'IQ0••', which means: interviewer (I) poses question (Q) from the questionnaire (0). A formal distinction can be made for the type of question (i.e., open, choice or yes-no questions, etc.) that is read. To indicate this, the fourth coding variable of our coding scheme (SPECIFICATION) is used. A code that specifies the question type as read by the interviewer may inform the researcher how the interviewer altered question wording. For example, a scripted closed question (e.g., 'How many bicycles do you own') receives the codes 'IQ0C•' (interviewer reads choice question). This question may actually be read as a yes-no question (e.g., 'Do you own more than one bicycle?') which then receives the code 'IQ0Y•', where the Y signifies that the question is read as a yes-no question. When a choice question is reworded as a yes-no question it is likely that the respondent gives 'yes' or 'no' as an answer. When yes and no are not predefined response alternatives, this is an inadequate (mismatch) answer. In only one other coding scheme (Loosveldt 1985) similar specification codes of question types were found.

The introductory text prior to a question (or a series of questions or assertions such as a 'battery') informs the respondent about the topic of the question(s) and may include some instructions (telling the respondent to use show cards, mentioning the response alternatives, offering definitions etc.).

The behavioral category 'instruction' means that the interviewer gives task-oriented information about the respondent's task. In our coding scheme a scripted instruction cannot

be distinguished from an introduction. Both receive the same code (i.e., 'IQ0I•', interviewer reads introduction from questionnaire). The decision to distinguish between these two scripted behaviors may depend on the specific survey that is analyzed. When a lot of different instructions and introductions are included in the interviewer's script, it may be useful to include specific categories to distinguish the two types of behaviors.

Evaluation of unconditional scripted behavior

Scripted behavior is usually evaluated for adequacy to investigate the influence of different types of inadequacy on the interaction and the quality of the response obtained. What behaviors are considered as 'adequate', 'inadequate' or 'neutral' with respect to interviewing rules depends on the researcher's own view (i.e., strictly or less strictly adhering to standardized interviewer instructions).

A common distinction of several coding options to evaluate adequacy of *question* reading is in terms of 'reading exactly as worded', 'reading with a minor change' and 'reading with a major change'. The code categories for 'exact' and 'minor change' may be combined in one code. This combined code category then indicates that the question was read in an acceptable/adequate way, though not necessarily exactly as worded.[8]

What is a major change in question reading is not always clear. Many types of inadequate reading are often included within the same category. The addition or deletion of a specific number of words or the fact that the meaning of a question is altered can be used as criterion to distinguish major changes from minor changes (Schaeffer and Dykema 2004), but also very detailed codes can be used to indicate whether words or phrases were added or deleted (Blair 1978).

In our coding scheme, we distinguish 'adequate', mismatch', 'suggestive', and 'invalid' questions. For example the question 'How many bicycles do you own?' can be read in various ways, as is illustrated in Table 4-3.

Table 4-3 The same question read in four different ways

Question read as:	Code	Interviewer reads question:
'How many bicycles do you own?'	IQ0CA	Adequately (no change in reading)
'How many bikes do you have?'	IQ0CM	Mismatch (meaning of question is not changed)
'You don't own any bikes, do you?'	IQ0YS	Yes-no instead of choice and Suggestively (interviewer suggests alternative)
'How many bicycles have you bought?'	IQ0CI	Invalidly (meaning of question is changed)

[8] Nolin and Chandler (1996) provide support for this strategy. In their study they found that, among the codes they used, the least agreement among coders seemed to exist between the codes for 'exact wording' and 'minor wording change'. This finding is also supported by levels of agreement for the code 'minor wording' in other studies. It appears that 4 of the 5 studies that present a Kappa value for this code, show a value below moderate agreement (i.e., below .50).

Exactly reading the *response alternatives* indicates that the interviewer has read the complete list of alternatives as worded in the questionnaire. In our scheme this is coded as 'IQ0AA' ('interviewer reads response alternatives adequately'). Whether changes in the reading of the response alternatives are inadequate depends on how strict standardization rules were applied. It might be adequate to use incomplete lists when the omission of alternatives is based upon earlier information of the respondent. When the respondent has made clear that one or two alternatives do not apply anyway, it may be even awkward to repeat all alternatives. However, according to rules of strict standardized interviewing all scripted response alternatives need to be read in all cases (Fowler and Mangione 1990).

Another frequently used evaluation of scripted behavior is adherence to questionnaire routing (i.e., omitting questions or reading of wrong questions). For example, interviewers may not verify information that respondents already provided in an earlier Q-A sequence. Because we use our coding scheme with the Sequence Viewer program, a skipped question can be identified by means of a sequence record lacking codes (i.e., there is no interaction).

4.3.2 Conditional (un) scripted interviewer behavior

Conditional behaviors occur after some kind of answer of the respondent. In this case the response (or lack of response) creates conditions for the interviewer's behavior. According to Oksenberg et al. (1991), coding interviewer behaviors after the initial asking of the question was "superfluous" since this behavior "tends to be reactive to respondent behavior" (Oksenberg et al. 1991, p. 352). With interaction analysis, we explicitly aim to study this reactivity, and this is the reason to include conditional behavior.

Scripted conditional interviewer behaviors comprise repetitions of questions and response alternatives. If sequential analysis is applied on fully coded data, it is possible to determine whether a question is the first or a repeated delivery, and how many times a question is read within the same Q-A sequence. Therefore it is not necessary to include information in the codes about the number of times questions have been read. Scripted conditional behaviors may also comprise scripted probes, and scripted clarifications or feedback. However, probes, clarifications and feedback are usually not scripted, so we usually consider those behaviors unscripted conditional behaviors.

Evaluation of conditional (un)scripted behavior

Scripted conditional behaviors can, like initial question reading, be evaluated in terms of adherence to scripts. A specific type of evaluation of both scripted and unscripted conditional behavior is the judgment of necessity or redundancy of the behavior. This judgment is often expressed in behavioral categories such as 'unnecessary probe/clarification' or 'fails to probe/clarify'. However, in many studies the meaning of necessity is unclear. Obviously, a probe, clarification or repetition of the question may be required when no adequate response from the respondent was received yet, but it is a difficult or even impossible judgment task to establish whether the behavior coded as absent, was the *only* appropriate behavior for an

interviewer to perform. Several behavioral options are available to interviewers instead, and it seems a bit biased to code for the necessity of specific behaviors. A better option is to use a general code 'fails to obtain an adequate answer from the respondent'. We will apply sequential analysis of fully coded data, in which case such general information may be deduced from the interaction.

Clarification

Clarification involves the interviewer giving information about the task of the respondent, or explanation of the question, by means of giving (scripted) definitions, or confirming the meaning of a question. In our coding scheme the code for interviewer's clarification is 'IQ0M•' (literal meaning: interviewer reads question clarifying its meaning). However, this only comprises unscripted clarifications (i.e., in interviewer's own wording). When interviewers repeat introductions or scripted instructions the code 'IQ0I•' (interviewer reads introduction from questionnaire) is used.

Evaluation of clarification

From a strict standardized point of view, any substantial clarification must be considered as inadequate. Nevertheless, a 'WIMTY'-response (i.e., 'Whatever it means to you', Moore, 2004) may be considered as the default-clarification from which interviewers may deviate in more or less adequate ways.

In our coding scheme we again use the concepts 'validity' and 'suggestiveness'. This yields three distinctions: adequate clarifications ('IQ0MA'), invalid clarifications (changing the question's meaning, 'IQ0MI') or suggestive clarifications (implying one or only part of the response options, 'IQ0MS'). Such distinctions may be useful in interaction analysis, e.g., to examine the causes of inadequate clarifications. We do not include detailed specifications (e.g., whether clarification expands or restricts meaning, see Loosveldt 1985). Such specifications may be useful when for instance effects of clarifications on response accuracy are studied. In response accuracy studies validating information is required, which is usually difficult to obtain for normal surveys.

Probing

Behaviors that are chiefly considered as 'probing' comprise the interviewer posing a question (usually more or less in her own wording), after respondents did not answer or gave an inadequate or incomplete answer. Probing can be prompted by the questionnaire either as a non-scripted probing instruction or as a scripted probe. Furthermore, the probe can be a request to the respondent to clarify the meaning of the answer or just to repeat the answer or to encourage the respondent to give further information.

Viterna and Maynard (2002) hold a broad definition of probing; all actions interviewers can perform when respondents did not give an adequate answer. They give three categories of probes: general probes (repetitions of questions, response alternatives, 'WIMTY'-responses, etc.), probes for closed questions (helping respondents to adjust their answer: "Which comes

closest…"), and probes for open questions (helping respondents to elaborate their answer "Could you tell me more about that?").

In our coding scheme no specific coding option is included to code for probes as such. Instead, several coding options are used to indicate the *function* and *format* of the probe. For example, interviewers can request respondents for clarification of their answer ('IR0m•': 'interviewer requests meaning of direct answer'), or interviewers can request for repetition ('duplication') of the answer ('IR0d•'). Furthermore, interviewers can repeat the question or alternatives (which both receive the same codes as initial question reading, i.e., 'IQ0CA' and 'IQ0AA' respectively).

Evaluation of probing

Probing can be evaluated for adequacy, e.g., resulting in code categories for non-suggestive or suggestive probes. In only a few coding schemes a specification is given for different types of suggestive probes (i.e., creating the distinctions 'leading', 'directive' and 'implied' probing, e.g., Smit 1995, Brenner, 1982 and Loosveldt, 1985). In our coding scheme, again the concepts 'mismatch', 'validity' and 'suggestiveness' are used. For example, interviewers may repeat the alternatives not precisely as worded ('IQ0AM'; mismatch alternatives), suggest a particular alternative ('IQ0AS'; suggestive alternatives), or repeat the question with a changed meaning ('IQ0CI'; interviewer reads invalid choice question).

Feedback

Feedback comprises interviewer behaviors that deal with acceptation or acknowledgement of the response. With these actions the interviewer in fact gives information about the adequacy of the response. Feedback usually occurs as a third-turn utterance, after the question and answer (Viterna and Maynard, 2002). In conversations, this third turn can indicate acknowledgments))and assessments (see section 2.2.10, chapter 2).

In our coding scheme two categories can be grouped under feedback. The first is perception. The interviewer can acknowledge a response by means of a simple perception such as 'mhm' or 'yes' ('IP0n•': 'interviewer perceives utterance by means of notification'), or by means of a repeat of the respondent's answer ('IP0E•': 'interviewer perceives utterance by means of echoing it'). The second category that can be grouped under feedback is a comment ('IC0t•': 'interviewer gives task related comment', like "That's useful information" or 'IC0p•': 'interviewer gives personal comment', such as "I can imagine that").

In their study, Bradburn and Sudman (1979) conclude that feedback was the most difficult behavior to recognize for coders. According to their estimation only just over one half of instances of feedback was coded. In their, and most other coding schemes, 'feedback' is used as an aggregated type of behavior, and as a consequence it is not always clear what specific actions a coder is supposed to include.

Evaluation of feedback

Feedback can of course be evaluated as adequate or inadequate. As Viterna and Maynard (2002) state, in the survey interview "fine graduations in these responses [i.e., feedback] potentially influence the respondent and create measurement error" (Viterna and Maynard, 2002:372).

Detailed codes for feedback are included in some schemes, such as 'positive' and 'negative' feedback (Blair 1978; Carton 1999; Loosveldt 1985), approvals for inadequate respondent behaviors (Lepkowski et al. 2000) 'long' versus 'short' feedback (e.g., Mathiowetz and Cannell, 1980, Lepkowski et al., 2000).

In our coding scheme, feedback behaviors can be derived from a coded Q-A sequence, for example: a neutral perception after inadequate respondent behavior indicates approval of such behavior, and the distinction between a simple perception ('IP0n') and a repetition of the respondent's answer ('IP0E') may give an indication of the length of feedback offered. Comments (task and person oriented) are not further evaluated for adequacy. Especially negative comments (e.g., 'You are a silly respondent' or 'I don't like my job as an interviewer') do not occur very frequently, and therefore the distinction between task and personal comments is informative enough.

Perceptions, more specifically: echoes of the other party, are evaluated for adequacy of the repetition with the concept 'mismatch' (i.e., yielding the code string 'IPOEA' for adequate repetitions and 'IP0EM' for mismatch repetitions). When an interviewer does not adequately repeat the respondent's answer, but instead provides a paraphrase not adequately representing the respondent's answer, this is considered an interpretation, which is expected to influence the eventual score negatively: respondents are not likely to correct a wrong interpretation.

4.3.3 Irrelevant interviewer behavior

Behaviors are considered as 'irrelevant' when they are not directly related to the questions of the questionnaire. A distinction can be made between 'detours' and utterances that indicate that the speaker is wandering off from the subject. Detours are temporary interruptions of the conversation (for example, the interviewer asks 'can I use your bathroom?'), and are not relevant here.

So-called 'wandering off' concerns utterances about topics that to some extent originate from topics of the questionnaire (i.e., elaborations). Wandering off is coded by means of the phases of the variable DISTANCE (i.e., '0' directly relates to the question, '1' indirectly relates to the question, and '2' concerns elaborations).

Other coding options that are sometimes included in coding schemes are related to the control of digression. For example, an interviewer can provoke a respondent into irrelevant behavior by asking irrelevant questions (coded as 'IQ2••' in our scheme). Failure to control digressions may be derived from fully coded Q-A sequences, during which a respondent continues to digress for several utterances.

4.3.4 Recording answers

Recording an answer, by entering it in the computer or writing it down is not a verbal action As a consequence, in only six coding schemes one or a few code categories are included related to recording answers, specifying mistakes an interviewer can make in recording answers. To this end, coders should have available the entered response.

To trace discrepancies between response and score, most researchers use only one code category indicating that such a discrepancy exists. Discrepancies can be accomplished by accident (a typing error) or intentionally. Van der Zouwen, Dijkstra and Smit (1991) give two possibilities of intentional discrepancies. The first is 'ignoring'; the respondent has produced a substantial response but the interviewer fills in 'don't know' or 'no response' on the questionnaire. The second is 'interpreting' (or 'choosing', see Van der Zouwen and Dijkstra 1995); the respondent does *not* choose one of the response alternatives, thus the interviewer chooses an alternative instead of the respondent.

In our coding scheme, the act of recording is not included as a separate coding option. Information about the score as entered by the interviewer is available to coders as a sequence variable (see section 4.2). This sequence variable may be compared with the actual answers given by respondents during the interaction (as indicated by the coding variable DIRECTION, see appendix 4-1).

4.4 Respondent behaviors

We found 78 different codes for respondent behavior in the 48 coding schemes. Table 4-4 gives an overview of behavioral categories that can be distinguished for respondent behavior, examples of specific behaviors, and evaluation criteria that are often applied in behavior coding schemes.

Table 4-4 Respondent behavioral categories and evaluation criteria

Type of respondent behavior	*Specifications*	*Evaluation criteria*
Answers	Adequate answers, Mismatch answers, Invalid answers, Qualified answers, Indirect answers	Codability, Understanding, Certainty, Accuracy, Relevance
Requests	Request for clarification, Request for repetition	Interpretations included
Non-answers	Don't know answers, Refusals	Assumed reasons
Feedback	Elaborations, Perceptions Comments	Informativeness, Relevance
Irrelevant behavior	Task related, Personal, Reactional to I or third person	-

4.4.1 Answers

The most often used distinction for answers is 'adequate answer' and 'inadequate answer'. For the latter the rather broad description 'does not meet question objective' is often used. In our coding scheme we use two types of inadequate answers, using the concepts 'validity' and 'mismatch'.

An invalid answer is coded with 'RA0AI' ('respondent gives answer by means of a choice of alternatives, but invalidly'). This is an answer that indicates misunderstanding of the question, as far as can be determined by the interaction. For example, a respondent may answer the question "Where did you watch television?" with "I watched the news". This answer indicates that the respondent apparently understood the question as "What (program) did you watch on television?" Such an answer requires the interviewer to clarify the question.

A mismatch answer (code 'RA0AM') is an answer that is not formatted according to the response alternatives, and needs probing by the interviewer. For example, a respondent may answer the question "How many days a week do you watch television?" with 'most days' instead of a number between 0 and 7. However, when the response evidently refers to one of the response alternatives, this is not considered a mismatch answer, because the interviewer is not required to probe. Such a response for the example of the television question may be 'never', which evidently refers to '0 days'. However, with 'every day' we do not know if the respondent refers to week or weekend days, so the interviewer at least needs to verify that.

This concept is related to Beatty's (2004) four levels of 'response precision', that differ with respect to the amount of rounding, judgment, or interpreting that is necessary for the interviewer to translate the response into an adequately formatted response. When answers are adequately formatted no rounding (and therefore no probing) is necessary, but inadequately formatted answers may vary in response precision.

Another distinction of answers is the concept 'qualification'. Respondents can qualify answers with respect to accuracy or certainty, e.g., with words like 'probably' or 'about' (see Dykema et al. 1997). Linguistic indicators of uncertainty are related to response accuracy (Draisma and Dijkstra 2004). Therefore, it may be useful to include the occurrence of doubt expressions in the codes for answers. In our coding scheme an option was included for qualified answers ('RA0AT': respondent gives answer by means of a choice of an alternative but with a qualification of uncertainty).

In addition, a respondent can think aloud while answering, or give a projective report (Moore (2004). In our coding scheme, reports (i.e., provisions of relevant information), think aloud utterances, and motivations are all coded with 'RA1O•'; respondent (R) provides an answer (A) that indirectly refers to the question (1) and is formatted as an 'open answer' (O), which we refer to as 'considerations'. Reports, thinking aloud, and motivations are difficult to distinguish from each other. These utterances often have in common that it is possible to infer an answer from the information provided. It is at least relevant to code these utterances with this summarizing code ('RA1O•'), as interviewers are likely to infer answers, and in that way influence the quality of the response obtained.

Finally, some distinctions between different types of answers we found in the coding schemes are redundant or even problematic to use in a coding scheme that is designed to analyze interactional patterns. Most of these specifications appear to depend on codes that precede or follow the behavior to be coded, thus making the interpretation of empirically found relations between different types of behavior difficult to interpret. Examples are: 'answers based on information of the interviewer' (Carton 1999, Loosveldt 1985), or 'answers that are incomplete or incorrect because the interviewer read the question incorrectly' (Marquis and Cannell 1969). Other categories or distinctions are not useful to include in our scheme because they may be derived from the Q-A sequence by means of sequential analysis. For example, 'additional response' (Sykes and Collins 1992), 'repeat of previous answer' (Loosveldt 1985, Marquis and Cannell 1969) or 'confirming a response already given' (Prüfer and Rexroth 1985).

4.4.2 Non-answers

Item non-response is traditionally divided into 'don't know' answers and explicit refusals to answer. In our coding scheme these two categories are included with the code strings 'RA0b•' (respondent answers 'don't know') and 'RA0r•' (respondent refuses to answer). 'Don't know' answers and refusals may indicate the difficulty of questions or sensitivity of question topics. Interviewers have two options of dealing with don't know answers or refusals: accepting them, or continue probing for a substantial answer.

'Don't know' answers can be considered as 'adequate' (i.e., 'codable') when they are listed as possible response alternatives. Some researchers include specific codes to account for (assumed!) reasons why a respondent cannot answer (Prüfer and Rexroth, 1985, Morton-Williams, 1979). However, respondents often do no express their reasons, and therefore this distinction may not be very useful.

4.4.3 Requests

A 'request for clarification' comprises the respondent asking for help from the interviewer. The request may concern clarification of the (meaning of the) question or response task ('RR0m•': respondent requests meaning of question), or a repetition of the question ('RR0d•': respondent requests 'duplication' of question). Some coding schemes include evaluations such as correct or incorrect interpretations incorporated in the respondents' requests. However, the specific issue of misunderstanding that is communicated with requests for clarification is very difficult to anticipate with specific codes. It is more useful to find out more about details in misunderstanding with a close look at the particular questions that frequently need clarification, or to inspect the actual transcriptions, than to include such specific codes.

4.4.4 Feedback

The behavioral category that is considered as 'feedback' comprises the respondent accepting or acknowledging the questions. In our scheme, feedback codes for the respondent comprise the same categories as those for the interviewer (i.e., perceptions and comments, see section 4.3.2). Feedback from the respondent is not often included in coding schemes, and with less detail as compared to interviewer feedback. This may be due to the fact that answers (especially the indirect ones) and feedback may be difficult to distinguish from each other. Both answers and feedback may provide information that is (indirectly) relevant to the questions, the interview or the survey.

Feedback is typically a 'summary' code (Blair 1978; Brenner 1982; Marquis and Cannell 1969), but might provide interesting deviations from the paradigmatic sequence. For example, when a respondent gives a lot of comments on the questionnaire, this may indicate task involvement (i.e., the motivation to approach the interview seriously). When many respondents independently criticize the same question this not only may indicate that respondents are generally involved in their task, but also that they may have problems with the question.

4.4.5 Irrelevant respondent behavior

The behavioral category 'irrelevant behavior' concerns behavior that is not directly related to the questions of the questionnaire. Like irrelevant behavior of the interviewer, a distinction can be made for 'detours' ('Wait there is someone at the door') and utterances that indicate that the speaker is wandering off from the subject. In our scheme, the stages in wandering off from a subject can again be coded with the variable DISTANCE (i.e., '0' directly relates to the question and '1' indirectly relates to the question). It is not possible to infer an answer from answers with a distance greater than 1, therefore these answers are considered as irrelevant.

4.4.6 Interruption

Interruption involves an event that concerns two utterances; the one being interrupted and the one that causes the interruption. Therefore, the coding schemes differ with respect to which actor the interruption is assigned to. In 19 coding schemes it is included as a code category 'respondent interrupts interviewer'. However, this can create difficulties for the coder, as the interruption is usually done by means of an utterance that comprises communication of some type of information; respondents interrupt the interviewer because they give answers, request for clarifications, provide feedback etc. Therefore, coding 'interruption' as a behavior on its own may complicate the scheme when only one code can be assigned to the utterance; it may be necessary to specify that the interruption itself was an answer or a request for clarification.

In two coding schemes (Sykes and Collins 1992, and Prüfer and Rexroth 1985) interruptions are coded by means of the utterance that was subject to the interruption (i.e., 'interruption of question reading'). In our coding scheme, interruption is included as a separate specification of the utterance that was interrupted, by means of the variable DIRECTION (see appendix 4-1). When respondents interrupt question reading, and especially

the alternatives, they are likely to cause mismatch answers, since they are not fully informed about the response alternatives due to their interruption.

4.5 Third party and general codes

Next to the interviewer and respondent, another person, i.e., a 'third person' (for example the partner or child of the respondent) may produce verbal behavior that can be coded with a certain level of detail. Behavior of third parties will influence the interaction and therefore is relevant to be coded. Two main categories of coding options may be relevant: the respondent or the interviewer talks to this third person (Marquis and Cannell 1969), or the third party talks to the respondent and/or interviewer. In our coding scheme, the coding options for third parties in principle comprise the *same* range of utterances that can be specified for respondent behavior.

Some coding schemes include coding options that refer to behavioral categories that cannot be attributed to the interviewer, respondent or third parties. Especially in coding schemes that apply full coding, coding options for missing data (i.e., unintelligible utterances) are necessary.

4.6 Conclusion

The existing coding variables, their possible values, and the hierarchal structure of the coding scheme generally proved to be appropriate for analysis of a wide range of survey interviews (see Dijkstra and Ongena forthcoming). Specific surveys or specific situations, or new theoretical insights may require adaptation of possible code combinations, additional categories within the coding variables, or even addition of complete variables. Fortunately, the coding scheme is flexible for such adaptations. However, rigorous adaptations, or adaptations only applied to some surveys analyzed, are at variance with the desire of data to be comparable across surveys, and to keep the coding scheme as simple as possible. Only small adaptations were implemented in the coding scheme for survey data analyzed in this thesis (see appendix 4-1).

In the next three chapters results of exploratory, non-experimental and experimental analyses will be presented of three different surveys that were coded with the multivariate coding scheme as described in this chapter.

5 Problematic deviations in question-answer sequences

5.1 Introduction

In this chapter we will describe the results of interaction analysis of an existing survey. An interaction is considered as an interdependent sequence of utterances of the interviewer and the respondent. Both may cause deviations from the paradigmatic sequence (see section 1.3, chapter 1).

A distinction can be made between problematic or non-problematic deviations. Problematic deviations are utterances that may have negative consequences for the accuracy of the response obtained, whereas from non-problematic deviations we do not necessarily expect such consequences. Non-problematic deviations are for example perceptive behaviors ('uh's), repetitions of questions, detours, comments and considerations. Problematic deviations by the respondent may signal problems in the cognitive processing of the question, or problems in the respondent's attitude towards the survey (e.g., lack of task involvement, lack of motivation etc.). Such problematic deviations require some action of the interviewer (i.e., explaining the response task, or just stimulating the respondent to provide an adequate answer). Interviewer problematic deviations are behaviors that are not performed according to interviewers' instructions, and are likely to affect the accuracy of the response. For some interviewer and respondent behaviors, consequences for the accuracy of the response obtained are uncertain, and may even be positive rather than negative. Such uncertain consequences concern for example requests for clarification; when the other party adequately reacts to such a request, a problem may be solved that is likely to enhance the quality of the data (i.e., the respondent is able to answer the question with a correct understanding of the question). However, requests may also indicate interactional problems (i.e., respondents may request for clarification because the interviewer read the question incorrectly). Thus, in our analysis a relatively broad range of behaviors is considered problematic.

5.2 Types of problematic deviations

Five different problematic deviations produced by respondents and nine problematic deviations produced by interviewers were distinguished. These are listed and briefly described in Table 5-1 and 5-2. In these tables, also the codes according to the multivariate coding scheme that we used are given.

Table 5-1 Problematic deviations by respondents

Problematic deviation	*Description*	*Action required by interviewer*
Mismatch answer (code: 'RA0AM')	Uncodable answers, not formatted according to prescribed alternatives	Probe for adequately formatted answer (e.g., repeat alternatives)
Invalid answer (code: 'RA0AI')	Does not answer the question within the intended meaning	Clarify or repeat question
Request for clarification (code: 'RR0mx')	Question about the meaning of the question	Clarify question
Irrelevant answers (code: 'RA2••')	The information given is not directly relevant to the question, DISTANCE (see section 4.2.2) is greater than 1	Clarify response task (e.g., repeat question)
'Don't know'-answers and refusals (codes: 'RA0b•'and 'RA0r•')	Respondent is not able or willing to provide information	Motivate R to think about answer or to give an answer

Table 5-2 Problematic deviations by interviewers

Problematic deviation	*Description*
Mismatch question (code: 'IQ0CM', 'IQ0AM')	Question or alternatives are not read exactly as scripted, but not changed in meaning
Invalid question (code: 'IQ0CI' or 'IQ0YI')	Question is read with a changed meaning
Suggestive probing (codes: 'IQ0AS' or 'IP0EM')	Suggestion of one or part of the alternatives; i.e., an explicit question to be confirmed by the respondent or an inference, not requiring confirmation.
Irrelevant question (code: 'IQ2••')	Information is asked that is not relevant to the question
Choosing (code: 'RA0AA' is *not* present)	In the Q-A sequence R does not give an adequate response alternative at all, but a response alternative is scored
Request for clarification (code: 'IR0m•')	Request to clarify the meaning of an answer
Incorrect scoring	Although an adequate answer was given, the answer category scored is different from the last answer given by the respondent
Omission of alternatives (code: 'IQ0AA' is *not* present)	In the Q-A sequence the interviewer does not present any of the response alternatives
Incorrectly skipped question	Question which should have been asked is not asked (i.e., Q-A sequence is empty)

Table 5-1 also indicates what action of the interviewer is required to solve the problematic deviation produced by the respondent. In case of irrelevant answers and don't know answers or refusals, this requirement is far less strict than in case of mismatch answers and invalid answers. When interviewers persist in probing for an adequate answer after 'don't know' answers, this may even have negative effects on the accuracy of the response. Respondents may truly not have the required information available, in which case providing a substantial response makes no sense. Moreover, the relation between the interviewer and respondent (i.e., rapport) may be disturbed, which will also have negative consequences for the answers to subsequent questions. However, 'don't know' answers and refusals may indicate difficulty or sensitivity of the question topic, and in some cases problems in understanding.

Four of the nine interviewer deviations (i.e., omission of alternatives, choosing, incorrectly skipped questions and incorrect scores) refer to the absence rather than the presence of a code category. The occurrence of these deviations can be derived from information of a completely coded Q-A sequence, and (in case of scoring) information from completed questionnaires. For example, omission of response alternatives in a Q-A sequence can easily be derived from a completely coded Q-A sequence. When the code 'IQ0A•' (interviewer presents alternatives from the questionnaire adequately) simply is not present among the codes in a Q-A sequence, we know that the interviewer did not present the response alternatives.

In order to reveal choosing, we first establish that the code 'RA0AA' (respondent provides adequate answer on the question from the questionnaire) is not present among the codes. When the interviewer has nevertheless entered a score, we can assume that the respondent did not give (and did also not confirm) an adequate answer, and consequently the interviewer performed 'choosing' behavior.

5.2.1 Consequences of problematic deviations

The relation between problematic deviations in interactions and the quality of the data is not always as clear as assumed. Although the occurrence of problematic deviations seems to lower the quality of the data that are produced by the interview, the occurrence of problematic deviations is not a sufficient indicator of quality. Most problematic deviations that are solved *adequately* probably do not make much difference with respect to the quality of the data compared to paradigmatic sequences or sequences with only non-problematic deviations. However, some problematic deviations, like suggestive probing, cannot be repaired once the damage is done. Once the interviewer has suggested a particular response, the respondent quite probably supposes this response is more appropriate, or in the rare case of nonconformist respondents, deliberately chooses *another* response alternative. Other problematic deviations, when adequately solved, may even *improve* the quality of the interview. A request to clarify the meaning of a question, for instance, makes clear that the respondent does not understand the question. When the problem in understanding is adequately solved, data quality may be improved (see Schober and Conrad 2002). In Q-A sequences without requests for clarification, respondents may not understand the question

either, but other deviations may give indications of problems in understanding (e.g., invalid answers, which indicate misunderstanding). However, from a paradigmatic Q-A sequence we do not know at all whether respondents understood the question correctly.

Several behavior coding studies have related the occurrence of problematic utterances and the quality of the data, often by using validating information (Belli, Lepkowski and Kabeto 2001; Dijkstra and Ongena forthcoming; Dykema et al. 1997; Loosveldt 1995; Smit, Dijkstra and Van der Zouwen 1997). Dykema et al. (1997), for example, found that interviewer's rewordings of questions increased response accuracy, whereas Smit et al. (1997) found that interviewer's suggestive probes decreased response accuracy. Dijkstra and Ongena (forthcoming) showed that the occurrence of problematic deviations was negatively related to the accuracy of responses.

5.2.2 Causes of problematic deviations

A particular problematic deviation can be caused by a preceding non-problematic deviation. For example a respondent's request for repetition may cause an interviewer to clarify the question inadequately (instead of repeating the question). A problematic deviation may also be caused by other problematic deviations. For example, if respondents give a mismatch (i.e., not precisely formatted) answer this may prompt interviewers to probe in a suggestive manner. Such interactional factors can be included in the analysis by comparing the different utterances before certain problematic deviations. Such an analysis requires sequential information in the data coded. A study that incorporated this type of analysis is Smit (1995). Smit's study showed that the problematic utterance 'suggestive probing' especially occurs after another fault in the interaction.

Our first research question (which is part of the third general research question presented in section 1.5, chapter 1) thus is:

Which deviations in the interaction cause interviewers and respondents to produce problematic deviations?

Problematic deviations may also occur irrespective of preceding deviations. Such non-interactional causes may be specific characteristics of the questions, the respondent, or the interviewer. Investigation of such factors does not necessarily require sequential information in the data coded, which makes this type of research much easier. Much more studies exist that explore these causes of problematic deviations than studies looking for interactional causes. For example, with respect to questions, studies showed that long questions as compared to shorter ones increased the chance of errors and variance in question reading (Bradburn and Sudman 1979) and increased the chance of requests for clarification and qualified answers from respondents (Cahalan et al. 1994). Furthermore, studies have shown that questions with show cards yield less problematic deviations than questions without such

cards (Prüfer and Rexroth 1985; Sykes and Collins 1992). Moreover, hypothetical or ambiguous questions and inadequate response alternatives increase the proportion of Q-A sequences with a problematic deviation (Van der Zouwen and Dijkstra 1995).

All that respondents have to do to contribute to the paradigmatic development of a Q-A sequence is providing an adequate answer. However, it is possible that respondents are not able to do this due to cognitive problems, or are not willing to put effort in providing adequate answers, due to motivational problems. Various studies have shown that older and lower educated respondents might have more cognitive difficulties in answering questions than younger and higher educated respondents, and as a result cause more problematic deviations in the interview (e.g. Bradburn and Sudman 1979; Cannell et al. 1968; Loosveldt 1994; Prüfer and Rexroth 1985). Loosveldt (1997) argues that age and education can be used as proxy indicators for cognitive ability and communicative skills that are necessary to answer survey questions adequately.

With respect to effects of interviewers, studies have shown that interviewers who know that the interview is recorded produce less problematic deviations (Cahalan et al. 1994). Interviewers who did not receive training are more likely to probe suggestively than trained interviewers (Loosveldt 1985), and interviewers who are trained in a socio-emotional style probe more often in a suggestive manner than interviewers trained in a formal style (Dijkstra et al. 1985). Furthermore, effects from the interviewer can partly depend on respondent characteristics, e.g., interviewers are more likely to read questions exactly as worded to female respondents than to male respondents (Gustavson-Miller et al. 1991).

Smit (1995) also found non-interactional causes of suggestive probing, e.g., the type of question and the interviewer. However, no relation was found between the occurrence of suggestive probing and specific characteristics of interviewer or respondent (e.g., their age, gender, etc.).

Interviewers that were involved in the survey we will study are fairly homogeneous with respect to demographic characteristics (such as their age, level of education, and gender). Therefore in this chapter the effects of particular interviewer characteristics will not be taken into account.

Our second research question is:
How are different types of problematic deviations related to characteristics of questions and respondents?

The questions will be answered by means of exploratory analyses of an existing (i.e., not experimentally manipulated) dataset, concerning a survey that can be considered representative for surveys conducted across universities and research institutions.

5.3 Exploratory study

5.3.1 The data

The data used for analysis comprised 211 CATI interviews that originate from a survey that was conducted in 1996, for a study on the effect of mass media advertising (Smit and Neyens 2000). From the survey, 38 relevant questions (e.g., closed choice questions in a format that is generally used in survey interviews) were selected to be coded. The questions included behavioral questions about watching television, attitude questions about television advertising, and general questions (i.e., respondent's age, level of education and the size of their household, see appendix 5-1).

Nine different interviewers, all female university students, aged between 18 and 26, interviewed the 211 respondents. These respondents differed from each other with respect to their age, level of education and gender, and can be considered representative for the Dutch population.

Following our recommendations described in section 3.6, the interviews were coded based upon transcribed interviews. In order to perform interaction analysis we applied full coding with preservation of sequential information. The selection of the 38 questions resulted in a file consisting of 7,635 Q-A sequences,[9] and comprised 41,847 utterances.

5.3.2 Occurrence of problematic deviations

In 3556 of the 7635 Q-A sequences one or more problematic deviations occur. Table 5-3 shows the percentage of Q-A sequences with a particular problematic deviation. For example, the 33.9% in the first row, means that in 33.9% of the Q-A sequences one or more mismatch answers occur. In 39.9% of the Q-A sequences the respondent produced one or more problematic deviations. Because it is very well possible that a respondent produced for example both a mismatch answer and a request for clarification, the percentages for individual respondent behaviors do not add up to 39.9%. The last column of the table shows the percentage of the 3556 Q-A sequences with a problematic deviation, in which the particular deviation occurs as the *first* problematic deviation, irrespective of who produced the deviation. Thus, the 65.3% in the first row means that in 65.3% of the Q-A sequences with a problematic deviation, the first problematic deviation was a mismatch answer, initiated by the respondent.

Of course, only one problematic deviation can occur as the first in a Q-A sequence, thus the total percentage in the last column equals the sum of the individual percentages.

[9] Not all questions were asked to all respondents, nor were all utterances in all sequences intelligible, and therefore less than 211*38 = 8,018 sequences were analyzed.

Table 5-3 Percentage of Q-A sequences with a particular problematic deviation

	Percentage of Q-A sequences with a deviation of all Q-A sequences	Percentages of first occurrence of Q-A sequences with any problematic deviation
Respondent behaviors		
Mismatch answer	33.9 %	65.3 %
Invalid answer	2.4 %	3.4 %
Request for clarification	3.5 %	5.9 %
Irrelevant answer	2.1 %	2.5 %
Don't know answer	3.4 %	5.7 %
Any problematic respondent behavior	39.9%	83 %
Interviewer behaviors		
Mismatch question	0.6 %	1.8 %
Invalid question	1.1 %	1.6 %
Request for clarification	0.6 %	0.7 %
Irrelevant question	0.5 %	0.0 %
Suggestive probe	16.5 %	10.9 %
Omission of alternatives	6.7 %	-
Incorrect score	1.7 %	1.9 %
Choosing	18.5 %	0 %
Incorrect skip	(n = 83)[10]	-
Any problematic interviewer behavior	25.7 %	17 %

It appears that mismatch answers occur most frequently, not only with respect to overall occurrence (33.9%), but especially as the first problematic deviation in a Q-A sequence (65.3%). Although both the interviewer and respondent produce quite a lot of problematic deviations, it appears that the respondent is mostly responsible for the *first* problematic deviation occurring in a sequence. In 83% of the sequences with a problematic deviation the first problematic deviation is uttered by the respondent, whereas in 17% of the sequences the interviewer is the first to produce a problematic deviation. This is remarkable, as the interviewer by default is the first actor in all Q-A sequences, and in this way the interviewer has most opportunities to cause the first problematic deviation in a Q-A sequence.

[10] Dividing the number of skipped questions by the number of 7635 Q-A sequences would not make any sense and it is therefore difficult to compare with other problematic events. The number of skipped questions can be divided by the number of questions that was supposed to be asked (i.e., 7635 + 83). However, during skipped-question QA-sequences no verbal utterances take place, which makes comparison with other problematic events difficult.

5.4 Causes of problematic deviations: preceding states

5.4.1 States within a Q-A sequence

The problematic deviations and the pattern of the paradigmatic Q-A sequence were used to distinguish five types of states of Q-A sequences. The first type is a paradigmatic state. A Q-A sequence is in such a state as long as its utterances correspond to successively: the interviewer poses the question adequately, the respondent immediately gives an adequate answer, and as a third utterance the interviewer may acknowledge the answer.

The second type is a non-problematic deviating state. A Q-A sequence is in such a state as soon as the sequence is no longer in a paradigmatic state because of the occurrence of a non-problematic deviation, i.e., any behavior not covered by tables 5-1 and 5-2. This deviation is assumed not to have negative consequences for the quality of the response obtained. For example, respondents may ask for repetition of the question, or may give a consideration before their answer.

In the remaining three types of states, a problematic deviation occurs. A Q-A sequence is in a problematic state at the moment the problematic deviation occurs. When the Q-A sequence is in a solved problematic state, this problematic deviation is solved. For example, a problematic deviation such as a mismatch answer is solved when the respondent finally gives an adequate answer. In unsolved problematic states, the problematic deviation is not solved, e.g., the respondent has not given an adequate answer yet. As we indicated in section 5.2, it is practically impossible to solve the problem of suggestive probing. Thus, sequences that contain suggestive probing always remain in an unsolved problematic state after the suggestive probe.

Within a Q-A sequence, multiple problematic and non-problematic deviations may occur sequentially (see Example 5-1).

Example 5-1 Description of a possible interaction sequence

	Verbal utterance	Code	Description of state
1.	I: Do you consider yourself too skinny, too fat or just good?	IQ0CA	Poses question adequately: no deviation, *Status:* paradigmatic
2.	R: Excuse me?	RR0d•	Requests repetition: non-problematic deviation, Status: non-problematic
3.	I: Skinny, fat or good?	IQ0AM	Read mismatch alternatives problematic deviation, *Status:* problematic
4.	R: My am not chubby	RA0AM	Mismatch answer: problematic deviation, *Status:* problematic
5.	I: uhuh	IP0n•	Notification of the answer: non-problematic deviation *Status:* unsolved problematic
6.	R: But I consider myself just good	RA0AA	Adequate answer: problem solving deviation, Status: solved problematic
7.	I: just good	IP0EA	Repeats answer adequately: non-problematic deviation *Status:* solved problematic

The first *non-problematic* deviation in Example 5-1 occurs in line 2 (respondent requests repetition). A request for repetition, although it requires action from the interviewer, is not a problematic deviation because it does not signal a problem with the meaning of the question. The first *problematic* deviation occurs in line 3: interviewer reads alternatives inadequately; instead of 'too skinny, too fat or just good' she reads 'skinny, fat or good'. Another problematic deviation occurs in line 4: respondent gives mismatch answer. The respondent does not format his answer according to the response alternatives. Utterance 5 is a non-problematic utterance. This utterance is not intended to solve the problems, and the previous problematic deviations still remain unsolved. The status of the sequence at that moment thus is 'unsolved problematic'. Utterance 6 (i.e., respondent gives adequate answer) is a non-problematic utterance that solves the problematic deviation of utterance 4; the respondent now formats his answer according to one of the response alternatives. The status of the sequence becomes 'solved problematic', and remains 'solved' with the last utterance in line 7.

In this way, the Q-A sequence is divided into subsequent states. In the example, the Q-A sequences starts in a paradigmatic state, which becomes a non-problematic deviating state, a problematic state, an unsolved problematic state and finally a solved state. To find causes of particular problematic deviations, we will investigate how often such problematic deviations are preceded by one of these five different states. We will focus only on the *first* instance of a particular type of problematic deviation for each actor (e.g., the first mismatch answer, the first invalid answer, the first suggestive probe, etc.) The second instance of the *same* problematic deviation is by definition preceded by the first instance. Thus, the first

instance affects the state of the sequence immediately preceding the second instance. Hence, taking only the first instance of a particular problematic deviation into account provides a less blurred picture of preceding states. By means of analyses of states preceding particular problematic deviations initiated by the respondent (section 5.4.2), respectively the interviewer (section 5.4.3), we will give an indication of the course of the interaction that precede such problematic deviations, and whether the interviewer or respondent seem to cause subsequent problematic deviations. In section 5.4.4 we will discuss in more detail which particular types of problematic respondent behaviors precede interviewer behavior.

5.4.2 Causes of problematic deviations initiated by respondents: preceding states.

In Table 5-4 the distribution of the immediately preceding states is shown for the first occurring problematic deviation by the respondent in a Q-A sequence. In the rows of the table the five different problematic deviations by respondents are listed. The columns of the table show the different states that the Q-A sequence was in just before the problematic deviation occurred. [11]

Table 5-4 Percentages of states immediately preceding problematic deviations initiated by respondents

	State of Q-A sequence immediately preceding problematic deviation by respondent						
First instance of:	Paradigmatic	Non-problematic	Solved problematic	Problematic	Unsolved problematic	Total	Number of sequences
Mismatch answer	75.8%	12.8%	2.7%	2.5%	6.2%	100%	2587
Invalid answer	34.9%	26.0%	7.1%	5.3%	16.7%	100%	180
Request for clarification	61.4%	15.4%	2.1%	7.2%	11.2%	100%	265
Irrelevant answer	12.1%	41.5%	5.3%	7.1%	35.1%	100%	159
Don't' know/ refusal	39.8%	30.8%	5.5%	7.2%	18.4%	100%	261

The first percentage for mismatch answers means that in 75.8% of the Q-A sequences with mismatch answer the state of the sequence was paradigmatic until the mismatch occurred. Thus, mismatch answers are typically preceded by a paradigmatic state, which indeed indicates that they occur spontaneously. In 12.8% of the cases they are preceded by a non-problematic state. Most of these non-problematic deviations were also initiated by respondents: they give considerations before their answer. Mismatch answers are not often preceded by problematic states due to other problematic deviations. Thus, mismatch answers

[11] It could have been informative to compare these percentages with a percentage based upon expected frequencies. However, it is rather dubious to compute expected frequencies of the states of Q-A sequences, because these are not independent. For example, once a deviation occurred, the sequence cannot return to a paradigmatic state. Even more problematic is that solved and unsolved states are by definition preceded by problematic deviations.

are unlikely to be preceded immediately by another problematic deviation (2.5%) or by another problematic deviation that was already solved (2.7%). An unsolved problematic state precedes mismatch answers more often (6.2%). The problematic deviations that accounted for such a state were mostly initiated by respondents as well: in half of the cases they first provided an invalid answer that was not solved (i.e., they did not give an adequate answer) before the mismatch answer occurred. The main conclusion for mismatch answers is that they occur spontaneously, and are not specifically caused by problematic behavior of the interviewer.

A large part of the invalid answers are preceded by a paradigmatic state (34.9%). However, invalid answers also are frequently preceded by a non-problematic state. Non-problematic deviations that cause such a state are again considerations; respondents first give an 'explanation' of their invalid answer (which may be a kind of spurious relation: this explanation enabled the coder to judge the answer as invalid; see also section 3.2.3). Furthermore, invalid answers are relatively often preceded by an unsolved problematic state. Invalid questions appear to partly account for this relation. In those cases interviewers cause respondents to answer invalidly, because the question was posed or explained invalidly. However, an invalid question does not necessarily mean that the subsequent answer is invalid. It is possible that respondents provide adequate answers, despite of the preceding invalid question, because the interviewer initially read the question adequately.

Requests for clarification also occur spontaneously, i.e., they are mainly preceded by a paradigmatic state. Nevertheless, in about 40% of the cases, requests for clarification are not preceded by a paradigmatic state, i.e., some deviation occurred before the request. However, we could not find particular deviations that typically account for these non-problematic or problematic states.

Irrelevant answers are more often preceded by a non-problematic state than by a paradigmatic state. Especially considerations (i.e., motivations and explanations of answers) often precede irrelevant answers. While the respondents are first explaining and illustrating, they are subsequently wandering off, the utterances becoming more and more irrelevant, and they do not seem to answer the question anymore. However, irrelevant answers are also relatively often preceded by an unsolved problematic state. This means that between the irrelevant answer and the preceding problematic deviations some other utterances occur, changing the state of the Q-A sequence into 'unsolved problematic'. Mostly, respondents gave a mismatch answer first, next the interviewer acknowledged this answer with some perceptive utterance, and then the respondent gave an irrelevant answer. So, also for irrelevant answers we can conclude that they are not often caused by problematic interviewer behavior.

Finally, 'Don't know' answers and refusals are mostly preceded by a paradigmatic state, but may also be preceded by a non-problematic state. Respondents provide a comment ('that is a difficult question') or first give a consideration ('I'd have to think about that') before they give a 'don't know' answer. No specific deviations can be identified that account for the problematic states that precede 'Don't know' answers and refusals.

In general, problematic deviations by respondents are more often preceded by paradigmatic states than by non-problematic and problematic states. The first occurrence of problematic deviations initiated by respondents are typically spontaneous events, without preceding problematic interviewer deviations. Apparently, the interviewer hardly ever causes respondents to produce the first problematic deviation in a Q-A sequence. In summary, we found for the interactional causes of problematic deviations initiated by respondents:

- Most deviations, especially mismatch answers and requests for clarification, occur spontaneously (i.e., without interactional causes)
- Invalid answers are typically preceded by a paradigmatic state, a non-problematic state (i.e., considerations), and to a less extent by an unsolved problematic state (i.e., invalid questions, which are the clearest instance of how interviewers can cause respondents to produce a problematic deviation)
- Irrelevant answers are often preceded by non-problematic deviations, which are mostly also considerations: i.e., the respondent is gradually wandering off the question topic;

5.4.3 Causes of problematic deviations initiated by interviewers: preceding states.

Because some of the problematic deviations of the interviewer concerned the absence rather than the presence of behavior, it is sometimes difficult to determine the place in the sequence where such deviations exactly occurred, and hence which state of the Q-A sequence preceded the deviation. This especially holds for omission of alternatives, because it is unclear when in the sequence the alternatives should have been read.

For skipping questions there is no interaction at all. Therefore, for these two behaviors it is not possible to detect interactional causes in the Q-A sequence, although these failures probably have such interactional causes. Choosing is a problematic deviation that also comprises absence of verbal behavior in the interaction, but we assume that the actual choosing, that is typing the response or selecting an alternative on the computer screen, occurred at the very end of the sequence (moreover, as soon as the interviewer typed the response, the CATI program automatically moved to the next question). Hence the state preceding choosing is the state at the end of the Q-A sequence. The same holds for incorrect scoring, which also should have been taken place at the end of the sequence.

Table 5-5 shows the distribution of the immediately preceding states for the first instance of each problematic deviation initiated by the interviewer in a Q-A sequence. In the rows of the table the seven different problematic deviations initiated by interviewers are listed. The columns of the table show the different states that the Q-A sequence was in just before the problematic deviation occurred. Because the interviewer is the first who produces an utterance in all Q-A sequences, we also included problematic deviations that were the first utterance of a Q-A sequence, i.e., not preceded by any other utterance. The ‘preceding’ state of the Q-A sequence is then considered paradigmatic; it is not preceded by some deviation.

Table 5-5 Percentages of states immediately preceding problematic deviations initiated by the interviewer

	State of Q-A sequence immediately preceding problematic deviation initiated by the interviewer						
First instance of:	Paradigmatic	Non-problematic	Solved problematic	Problematic	Unsolved problematic	Total	Number of sequences
Mismatch question	28%	8%	2%	57%	5%	100%	47
Invalid question	12%	8%	13%	53%	14%	100%	88
Suggestive probing	11%	16%	2%	57%	14%	100%	1261
Irrelevant question	0%	0%	0%	100%	0%	100%	44
Choosing	0%	0%	9%	16%	75%	100%	1408
Request for clarification	32%	19%	4%	32%	13%	100%	47
Incorrect scoring	13%	15%	8%	14%	50%	100%	131

Mismatch questions are in 28% of the cases preceded by a paradigmatic state. In many cases, this means that the interviewer starts the Q-A sequence with the mismatch question. However, mismatch questions are most often preceded by one of the problematic states (i.e., 2% + 57% + 5%), thus other problematic deviations occurred before the mismatch question. These occurrences refer to interviewer's repetitions of the question or response alternatives later on in the Q-A sequence, i.e., after problematic deviations. Typically, mismatch questions occur immediately after another problematic deviation (57%).

Also invalid questions and suggestive probing are usually preceded by (solved and unsolved) problematic states. Irrelevant questions and choosing are never preceded by paradigmatic or non-problematic states, but always by some kind of problematic deviation.

Requests for clarification initiated by the interviewer are equally often preceded by a paradigmatic as by a problematic state (i.e., both 32%). Requests for clarification appear to occur mostly when the respondent has given an adequate answer but the interviewer has doubts about the correct understanding of this answer.

In summary, problematic deviations initiated by interviewers are most often preceded by problematic states. In the next section we will discuss which particular problematic deviations (which turned the state of the Q-A sequence into a problematic one) cause the interviewer to produce her own problematic deviations. It seems plausible that the deficient behavior of the interviewer is not only caused by more or less immediately preceding problematic behavior of the respondent, but also by multiple earlier deviations. For example, respondents may request for clarification, and subsequently give a mismatch answer. After these problematic deviations, interviewers may clarify the question incorrectly (i.e., 'invalid question'). The cause of this last problematic deviation can be found both in the respondents' request for clarification and the mismatch answer.

5.4.4 Specific utterances preceding problematic deviations initiated by interviewers

In Table 5-6 the specific behaviors are shown that accounted for the 'solved problematic', 'problematic' or 'unsolved problematic' states that preceded problematic deviations by the interviewer. We counted the number of problematic deviations produced by respondents that preceded the first problematic deviation initiated by the interviewer. The analysis was done for all preceding utterances in the same Q-A sequence; in order to investigate whether, next to the immediately preceding ones, earlier deviations in the Q-A sequence also provide explanations for the occurrence of problematic deviations initiated by interviewers.

Table 5-6 Percentage of problematic deviations by interviewers preceded by specific problematic deviations by respondents

	Type of problematic deviation preceding problematic deviation of interviewer					
First instance of:	Mismatch answers	Invalid answers	Requests for clarification	Irrelevant answer	Don't know answer/ refusal	Total number of deviations
Mismatch question	94%	3%	3%	0%	0%	30
Invalid question	38%	8%	25%	9%	8%	70
Suggestive probing	89%	6%	6%	6%	8%	921
Irrelevant question	36%	7%	7%	100%	9%	44
Choosing	88%	6%	4%	6%	2%	1408
Request for clarification	92%	8%	18%	4%	0%	23
Incorrect scoring	29%	2%	3%	3%	36%	94

Because more than one problematic deviation initiated by the respondent may occur before a problematic deviation of the interviewer, the percentages in the rows can add up to more than 100%. For example, irrelevant questions are in all cases (100%) preceded by irrelevant answers, but in the same sequence, in 36% of the cases also preceded by mismatch answers.

Because in a Q-A sequence the problematic behavior of the interviewer may not be preceded by any problematic respondent behavior at all (but instead by other problematic interviewer behavior), percentages may also add up to less than 100%.

Problematic deviations by interviewers, especially mismatch questions, suggestive probing, choosing, and requests for clarification are mainly preceded by mismatch answers. Furthermore, suggestive probing and choosing are the most frequently occurring problematic deviations. Thus, the most important cause of problematic interviewer behavior is the occurrence of mismatch answers.

Invalid questions are fairly often preceded by respondent's requests for clarification. These invalid questions obviously concern invalid clarifications of the meaning of a question.

Irrelevant questions are in all cases preceded by irrelevant answers. Thus, interviewers do not evoke digressions, but once respondents start to digress they may stimulate them to continue digressing. However, an additional analysis, concerning the utterances that follow

irrelevant answers, showed that interviewers do not ask irrelevant questions after all cases of irrelevant answers (they do so in 27% of the cases).

Incorrect scoring is often preceded by 'don't know' answers. Although the 'don't know' alternative is available among the response alternatives, the interviewer nevertheless does not always score this option. An additional analysis, concerning the scores that were given in Q-A sequences with 'don't know answers', showed that interviewers score something else in 21% of the cases that respondents gave 'don't know' answers. Interviewers choose the middle alternative ('neutral' or 'neither agree nor disagree') to be scored instead of 'don't know' answers for 88% of these incorrectly scored 'don't know' answers.

In general, problematic deviations by interviewers are more often preceded by problematic states than by paradigmatic or non-problematic states. Interviewers do not spontaneously produce problematic deviations, but are mostly triggered by problematic behavior of respondents.

In summary, we found for the problematic deviations by interviewers:

- Mismatch answers occur most frequently, and are the main cause of problematic deviations by interviewers;
- Invalid questions are also preceded by respondents' requests for clarification;
- Irrelevant questions are always preceded by irrelevant answers, but in some cases also by other problematic deviations initiated by respondents;
- Incorrect scoring is often preceded by don't know answers.

5.5 Causes of problematic deviations: the questions

The 38 questions that were asked in the survey are likely to differ in the number of problematic deviations they evoke. Mismatch answers occur most frequently of all problematic deviations, usually were the first problematic deviation in a Q-A sequence, and the main cause of problematic deviations by the interviewer. Furthermore, questions differ mostly from each other on the occurrence of mismatch answers. Therefore, we will discuss only the occurrence of mismatch answers in our comparison of questions.

In Table 5-7 the percentage of sequences with a mismatch answer for each question is shown. The table shows that questions differ substantially from each other with respect to the occurrence of mismatch answers; the percentage of occurrence varies from 2% to 59% of the Q-A sequences (Cramer's $V = 0.39$, $p < 0.01$).

The difference between questions might be attributed to several characteristics of questions, such as the types of response alternatives used, the manner of presentation of alternatives, and the question order in the questionnaire. These will be discussed in the next sections.

Table 5-7 Percentage of sequences with a mismatch answer per question

	Percentage of Q-A sequences with a mismatch answer
1. Number of days watching TV	21%
2. Number of minutes TV per day	24%
3. Last time watched TV	6%
5. Watch out of interest or pastime	22%
6. Where watching TV	6%
7. Watching alone/with others	2%
8. Number of minutes watched	19%
9. Percentage watched with attention	42%
10. Number of commercial blocks seen	23%
11. Number of commercials seen	20%
12. Percentage commercials with attention	4%
13. Stay to watch (when commercials show up)	49%
14. Switch other channel (when commercials show up)	54%
15. Volume off (when commercials show up)	55%
16. TV-set off (when commercials show up)	56%
17. Do something else (when commercials show up)	59%
18. Leave the room (when commercials show up)	50%
19. Search for commercials	58%
20. Commercials comprise special offers	53%
21. Commercials are funny	56%
22. Commercials show other consumers	44%
23. Commercials show new products	43%
24. Commercials are entertaining	59%
25. Commercials show up at inconvenient moments	53%
26.commercials are too blaring	50%
27. Commercials are implausible	47%
28.commercials are repeated too often	46%
29. Commercials are too much alike	41%
30. Pay attention to commercials	35%
31. Positive/negative attitude TV ads	19%
32. Positive/negative attitude radio ads	23%
33. Positive/negative attitude newspaper ads	28%
34. Positive/negative attitude magazine ads	20%
35. Positive/negative attitude ads in general	21%
44. Number of persons house	2%
45. Age	4%
46. Employed	5%
49. Education	26%

5.5.1 Response alternatives

A characteristic of questions that might be related to the occurrence of problematic deviations is the type of response alternatives used. We used the type of alternative, and the manner of presentation of alternatives as criteria to create five question categories.

As is shown in Table 5-8, questions with implicit alternatives require response alternatives with a well-known meaning (such as numbers, a location, or a specific day of the week), and the interviewer is not required to mention the alternatives. Mentioning all alternatives after reading the question would be even awkward (e.g., in case of Q3,

concerning the day respondents watched television the last time, all days of the week would have to be read).

A substantial difference between the question types entails the way response alternatives must be presented. Questions with implicit alternatives, field coded questions and yes-no questions do not require the interviewer to read the response alternatives. Assertions do require reading of response alternatives, and this is done before the question or in a two-step procedure (see sections 5.5.2 and 5.5.3).

The table also shows that the percentage of Q-A sequences with a mismatch answer differs for the five question types.

Table 5-8 Five question-types for the 38 questions from the CATI survey

Type	Description	Questions	Percentage of sequences with a mismatch answer
Implicit	Closed question with *implicit* response alternatives (number, day of the week)	Q1-Q5, Q8-Q12, Q44-Q45	19%
Field code	Field coded question, i.e., with *listed* response alternatives, not read by the interviewer ('where', 'which')	Q6, Q49	17%
Assertion: alternatives before Q	Assertion with four response alternatives, read before the question by the interviewer ('each time-often-sometimes-never')	Q14-19, Q30	52%
Assertion: alternatives in 2 steps	Assertion with 5-point scale, response alternatives read in two steps (agree-disagree/positive-negative)	Q20-Q29, Q31-Q35	40%
Yes-No	Yes-no question	Q7, Q46	3%

Yes-no questions evoke the smallest number of mismatch answers. This effect also showed up in a number of different surveys (see Dijkstra and Ongena forthcoming). Questions with implicit alternatives and field-coded questions evoke a moderate number of mismatch answers, and assertions induce the highest percentage of mismatch answers.

For questions with implicit alternatives, respondents are asked to formulate their own answer within the prescribed range (i.e., a number of years, days, or a percentage). The questions adequately imply this answering format (e.g., 'how many minutes...', 'which percentage…'), and thus it might be relatively easy for respondents to format their own answer instead of picking an alternative from a list. Some of these questions can be considered difficult; such as the ones that ask the percentage of time respondents watched television (Q9) or commercials (Q12) with attention. Thus these questions yield the highest number of mismatch answers within this question type.

Assertions evoke *most* mismatch answers. These are questions with response alternatives that consist of a scale with an ordinal gradation (e.g., 'each time-often-sometimes-never') or a Likert-type scale ('strongly agree-agree-neither agree nor disagree-disagree-strongly disagree').

Respondents are not accustomed to the use of response alternatives created by researchers. Thus, they have more difficulty with the use of these kinds of words. The response alternatives for opinion assertions and factual assertions are presented in different ways; these strategies will be described in the next two sections.

5.5.2 Presentation of response alternatives in two steps

For opinion assertions the response task was intended to be simplified with a presentation of the response alternatives in two steps. For the first assertion of the battery, interviewers were required to read an introduction text that informed respondents about the complete set of alternatives. Next, an assertion was presented with two categories ('do you agree or disagree?'). After the respondent's first response that indicated the direction of the opinion, the interviewer had to read the extreme and moderate alternative of the chosen direction ('strongly agree or just agree?'). This in fact transformed the question into two subsequent questions, each with two alternatives, but this strategy created two problems.

A first problem is that interviewers often omitted the second part of the question (asking the intensity of the opinion, i.e., 'omission of alternatives'). The second (and more important) problem is that respondents often already replied with a mismatch answer to the first part of the question.

It is surprising that this initial question, that requires a choice between only two alternatives ('agree' and 'disagree') still yields a high number of mismatch answers, whereas yes-or-no questions, that also offer the choice of two alternatives, yield far less mismatch answers. Apparently, it is not the low number of response alternatives (i.e., two) that causes yes-no questions to be less problematic. Yes-no questions could be less problematic because the response alternatives 'yes' and 'no' are commonly used in ordinary conversations, and therefore easy to use for respondents. Categories like 'agree' and 'disagree' are formulated by researchers and not so frequently used in ordinary conversations.

Another explanation for the frequent occurrence of mismatch answers for opinion assertions could be that respondents are not reminded of the middle alternative ('neither agree nor disagree'), as this alternative is only offered at the introduction of the battery of assertions.

The lack of attention to the middle alternative also confuses interviewers. They do not know the difference between 'don't know' and 'neutral' and certainly are not convinced of the importance of this difference. The same two-step procedure was used for questions with '(strongly) positive', '(strongly) negative' and 'neutral' as response alternatives. The utterance in line 3 of Excerpt 5-1 shows that interviewers may even insist on answers that do not refer to the middle point of a scale.

Excerpt 5-1 Q-A sequence concerning Q33

1. I: Are you in general negative or positive towards newspaper advertising?
2. R: Yes make it neutral as well because...
3. I: Yes? But couldn't you eh because it is a little.. yes it is quite.. couldn't you say negative or positive or somewhat negative somewhat positive is also possible anyway
4. R: Yes make it somewhat
5. I: Yes somewhat negative or somewhat positive?
6. R: Yes somewhat positive

5.5.3 Instructions for repeating alternatives.

In the battery of assertions with four response alternatives (Q13-19, concerning respondent's behavior when commercials are shown on television), the same response alternatives are used for seven questions in a row. The introductory question (Q13), was preceded by a statement including all alternatives (i.e., "I will mention some possible reactions to television advertising. Would you please tell me whether you do this 'each time', 'often', 'sometimes' or 'never' when commercials appear on the screen?") . Each question asked respondents to respond to an assertion, like 'You stay to watch the commercials'.

Interviewers were implicitly instructed to repeat the complete set of alternatives for each subsequent question. This implicit instruction, did not literally tell them *how* to repeat the response alternatives; all that was available to them on the CATI screen was the assertion and below this assertion the list of alternatives with a reminding instruction text "read the alternatives". Therefore, not repeating the alternatives for the factual assertions was not considered a problematic deviation.

Interviewers rarely repeated the alternatives for each assertion. Apparently, interviewers expect (and probably often experience) that respondents will remember the complete set of alternatives from the first question.

According to a conversational view of interviewing (Suchman and Jordan, 1990), reading the alternatives for each and every question within the series could in fact be considered a violation of the maxime of quantity (i.e., give only new information, Grice 1975); the response alternatives are 'given' information (i.e., at the introduction of Q13), which need not be repeated for Q14-19.

Interviewers may have been prone to first read the assertion, and then optionally repeat the alternatives. Interruptions are very likely to occur when alternatives are read *after* instead of *before* or *within* the question delivery component (see section 2.2.9). Apparently, interviewers anticipated this likelihood of interruption by not reading the alternatives and instead awaiting the respondent's first response.

Not reading the alternatives is significantly related to the percentage of mismatch answers occurring. When the alternatives are read, the percentage of mismatch answers is lower (37%) than when they are not read (57%, $p < 0.01$, Cramer's $V = 0.11$). Apparently not reading the response alternatives, although appropriate from a conversational point of view, seems to increase the chance that a mismatch answer will occur. Although from the moment

of the introduction of the alternatives only four different response alternatives had to be remembered, for the seven subsequent questions with the same alternatives the respondents evidently need help from the interviewer. However, the fact that interviewers rarely read the alternatives shows that the instruction how to read the alternatives was not adequate.

5.5.4 Question order: general and specific questions

An aspect related to the order of questions, is the arrangement of general and specific questions or 'part-whole combinations' of questions (Schuman and Presser, 1996) in a questionnaire. According to Schwarz (1995), a general question following specific questions can be interpreted in two ways. Firstly, one can interpret the general question as a request to summarize the previous aspects in a general judgment. Secondly, one can interpret the general question as a new question excluding specific aspects that were asked before. The second interpretation can be illustrated by the maxime of quantity (Grice 1975) that evokes respondents to give only *new* information.

In the questionnaire, there is one example of a general question following several specific questions. This is Q35, a question concerning an overall judgment of the respondent's attitude towards advertising in general, that follows a series of questions (Q31-Q34about attitudes towards advertising on specific media (television, radio, newspapers and magazines). Q35 could be considered as a summarizing question, when aspects covered in Q31-Q34 are interpreted as exhaustive. However, the attitude towards advertising in general may also be affected by media not mentioned in the survey: e.g., advertising on billboards, on the Internet etc.

The way in which respondents will interpret the question depends on the extent to which the previous questions were exhaustive. However, the respondent is not the only one who needs to interpret the questions: interviewers must (and will) interpret questions as well. When interviewers interpret the general question as a summarizing question they do not expect new information from respondents. Therefore, they may decide to skip the question or only verify their expected answer with a suggestive probe. This is especially likely to occur when all answers of a respondent to the specific questions are exactly the same (e.g., 'disagree') and therefore the interviewer assumes the answer to the general question is 'disagree' as well.

It appears that 19% of the respondents reply with the same answer for each of the specific questions. In Table 5-9 the percentage of Q-A sequences with a problematic deviation initiated by interviewers with respect to Q35 are presented. In the first column the percentages are shown for respondents who replied with the *same* answer for all four preceding questions Q31-Q34 (n=40). In the second column the percentages are shown for respondents who replied with *different* answers for the preceding questions Q31-Q34 (n=171). Two statistically significant differences were found. The first is that interviewers skip Q35 (but nevertheless fill in some score) relatively more often when the previous questions were all answered in the same way (13%) than when they are not (2%). The second is that interviewers omit reading of alternatives relatively more often with Q35 when the

previous questions were all answered in the same way than when they are not. This indicates that interviewers find themselves better able to guess the answer of a respondent when the previous answers were all answered in the same way.

Table 5-9 Percentage of problematic deviations for Q35 when for Q31-Q34 the *same* or *different* response alternatives were chosen.

	Q31-Q34 *same answers*	Q31-Q34 *different answers*	Total	Cramer's V
Problematic deviations for				
Interviewer:				
Skipping question Q35	13%	2%	4%	0.20*
Suggestive probing	24%	16%	17%	n.s.
Omission of alternatives	40%	23%	27%	0.15*
Choosing	3%	14%	12%	n.s.
Number of sequences	40 (19%)	171 (81%)	211	-

To summarize, differences between questions in the percentage of problematic deviations are mostly related to the types of alternatives used, and the way they are presented to respondents. Overall, for questions as causes of problematic deviations we found:

- Yes-no questions, with appropriate alternatives, yield a low number of mismatch answers;
- Presenting a Likert-type scale in two steps creates problems for the middle alternatives, causing mismatch answers
- Inadequate instructions to repeat alternatives yield a high number of mismatch answers;
- General questions after specific questions cause interviewers to skip the general question or omit reading of alternatives.

5.6 Causes of problematic deviations: the respondents

Respondents can vary with respect to the number of problematic deviations they produced, but may also differ in the extent to which they trigger interviewers to produce problematic deviations. Variables that are available to specify such differences are respondents' age, level of education and gender.

Table 5-10 shows percentages of Q-A sequences with problematic deviations by respondents and interviewers for several relevant respondent variables. Problematic deviations are related to all respondent variables. Generally, older respondents produce problematic deviations in a Q-A sequence relatively more often (i.e., in 57% of the Q-A sequences for the eldest respondents) than younger respondents (i.e., in 25% of the Q-A sequences for the youngest respondents). Furthermore, the level of education of respondents has significant effects on the occurrence of problematic deviations. Less educated respondents produce relatively more problematic deviations (i.e., in 57% of the Q-A sequences for the lowest level of education) than higher educated respondents (i.e., in 29% of

the Q-A sequences for the highest level of education). When the effect of age is controlled for the level of education, the effect of age still remains significant. However, when the effect of education is controlled for the age of respondents, it appears that the respondents in the oldest age category (71 years and older) with different levels of education do not differ with respect to the percentage of Q-A sequences with a problematic deviation.

Gender of respondents also had a significant effect. Female respondents relatively more often produce problematic deviations than male respondents. However, when controlled for age, this effect is only significant in the oldest age category (65 and up). When controlled for education, the effect is only significant in the lowest educational level (i.e., primary education). Nevertheless, all non-significant differences between female and male respondents are in the same direction across age and education groups; i.e., female respondents produce more problematic deviations than male ones.

Table 5-10 Relation between problematic deviations and respondents

Respondent variable	Percentage of problematic deviations by respondents	Percentage of problematic deviations by interviewers	N
Total	40%	26%	211
Age:			
25 years and younger	25%	20%	32
26- 40 years	30%	22%	65
41-55 years	45%	27%	55
56-70 years	54%	30%	57
71 years and older	57%	36%	21
Unknown age	40%	17%	1
χ^2 =	360.52**	90.12**	
Education			
Primary education	57%	29%	12
Lower vocational	48%	28%	34
Secondary education	42%	31%	25
Vocational education	39%	25%	45
Pre-university education	31%	21%	28
Higher education	37%	25%	39
University	29%	18%	20
Unknown	52%	35%	8
χ^2 =	182.80**	69.23**	
Gender			
Male	42%	27%	98
Female	58%	73%	113
χ^2 =	17.74**	12.89**	

Problematic deviations initiated by interviewers are mainly caused by other problematic deviations, especially by mismatch answers (see section 5.4.3). The respondent variables age, gender and education are not likely to have a direct effect on problematic deviations initiated by interviewers. However, some direct relations could exist. Interviewers omit alternatives

(i.e., to ask for the intensity of opinions) more often when respondents are female, relatively young, and when they are highly educated than when they are male, relatively old, or lower educated. The former respondents mostly resemble the interviewer with respect to age and gender. Apparently, interviewers find themselves able to guess the intensity of an opinion when respondents resemble themselves with respect to their age, and thus interviewers do not suppose probing is necessary. It is also possible that interviewers are more likely to forget about standardized interviewing rules when they are talking to respondents of their own age and gender.

5.7 Summary and conclusion

In almost 50% of the sequences a problematic deviation occurs. The first problematic deviation is generally produced by the respondent. It appears that problematic deviations initiated by respondents are particularly caused by characteristics of questions and respondents, and generally do not have interactional causes. The mismatch answer is the most frequently occurring problematic deviation. A summary of the main results is visualized in Figure 5-1.

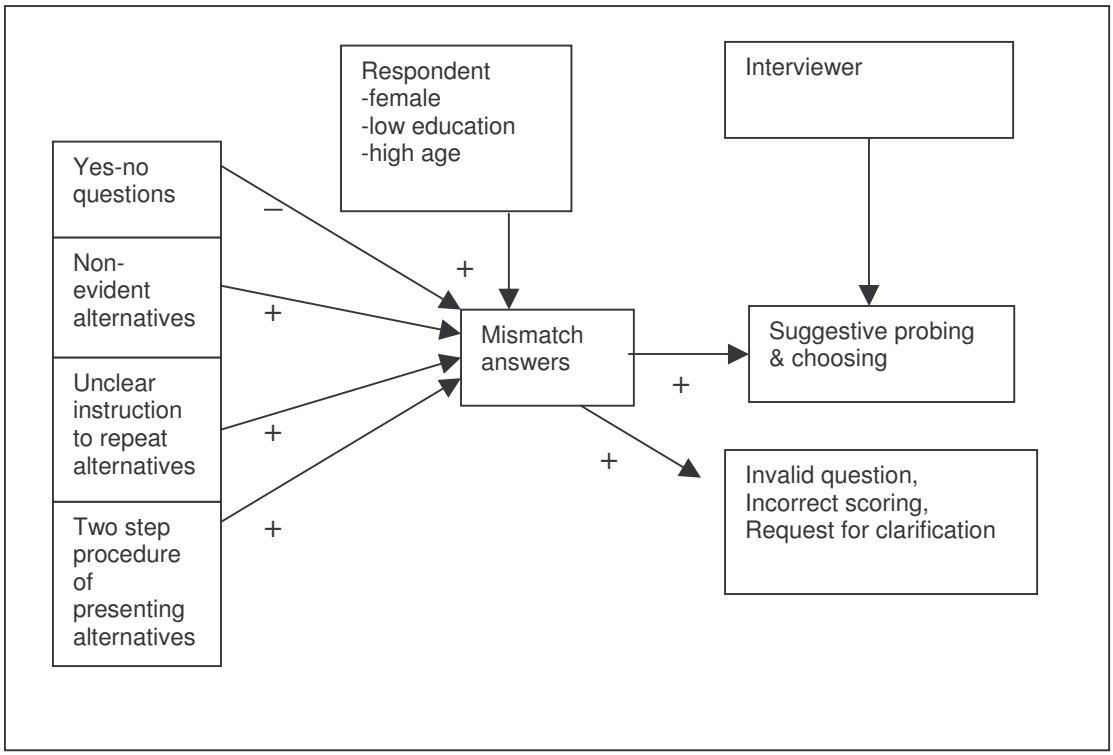

Figure 5-1 Summary of the main results

Several characteristics of the questions are related to the occurrence of problematic deviations. Questions with a list of response alternatives formulated by the researcher evoke more mismatch answers than questions with implicit alternatives. Especially yes-no questions yield a low number of mismatch answers. Respondents also differ in the number of problematic deviations they produce. Less educated respondents initiate problematic

deviations more often than higher educated respondents, and older more than younger respondents. Female respondents tend to produce problematic deviations relatively more often than male respondents.

Problematic deviations initiated by interviewers are primarily caused by interactional factors, i.e., preceding problematic deviations. Suggestive probing, omission of alternatives and choosing are the most frequently occurring problematic deviations initiated by interviewers. Of these three deviations, suggestive probing and choosing are mainly caused by mismatch answers. This finding corresponds to that of Smit (1995).

Interviewers are driven into suggestive probing because they have to verify an unclear answer. Mismatch answers almost always give clues for the response alternative that may be appropriate, and when interviewers assume they know the answer, they only have to verify it. Mentioning all other alternatives would appear to be strict, non-collaborative behavior.

It may be more effective to prevent occurrence of problematic deviations than to train interviewers how to handle problematic deviations initiated by respondents more adequately. Several practical solutions could be derived from the analyses discussed in this chapter. For example, response alternatives need to be read for each individual question, and not only during the first introductory question of a series of questions with the same range of alternatives. Although this is contrary to conversational principles, it appears that when the interviewer does not repeat the response alternatives, more mismatch answers occur. However, it is important that the interviewers are properly instructed when to present the alternatives. To accomplish such instructions, the alternatives should be incorporated in the question, i.e., they should be read before the question delivery component (from which the question meaning may be derived) is finished. When alternatives are incorporated in the question, interviewers are less likely to be interrupted by respondents, because the question delivery component is not known before the response alternatives are read. Respondents then are more likely to await the complete reading of the question, and thus are fully informed about the complete range of alternatives. As a consequence, they are less likely to give mismatch answers.

A practical solution to deal with long lists of response alternatives was the two-step procedure of presenting the response alternatives of the Likert-type scale for assertions. Although this seems to be a feasible solution, it appears that due to this procedure the middle alternative is only mentioned at the introduction of the battery of assertions. This not only creates a bias, it also seems to lead to a higher number of mismatch answers, partly because respondents are not reminded of the exact wording of the middle alternative. However, probably not all mismatch answers are due to respondents who want to choose the middle alternatives. The most likely cause of mismatch answers is the difficulty respondents have with the use of the response alternatives as created by the researcher.

From our results, we can conclude that the wording of questions and response alternatives seems to be an important explaining variable of the occurrence of mismatch answers. This may be viewed as an encouraging result, because question wording can usually

be directly controlled by researchers, which is less true for interviewer or respondent characteristics.

However, improving question wording based upon coincidentally found effects may not be very useful. What is needed is a theoretical explanation of the occurrence of mismatch answers. To find this explanation we can go back to the theoretical approaches that were described in chapter 2. Cognitive and conversational factors may account for the occurrence of mismatch answers. Mismatch answers may be an indication of problems in performing the cognitive steps of the Tourangeau et al. (2000) model (see section 2.3, chapter 2). Especially retrieval of information may be related to the occurrence of mismatch answers. Respondents may express retrieval of information through verbal considerations. This may especially be the case when respondents need to process a lot of information, for example enumeration of instances of frequent behavior during a relatively long period. Interviewers, who expect a response from the respondent, may even evoke such verbal considerations, when they stimulate respondents to provide information.

With respect to conversational factors, the problem is more a motivational one. Respondents may view the survey interview as an ordinary conversation, and as a consequence they may be convinced that conversational answers are adequate. This explains why respondents produce least mismatch answers when they are asked yes-no questions; such questions are perfectly normal in ordinary conversations and evoke a conversational response, i.e., a ‘yes’ or a ‘no’. Questions that are formulated as assertions also trigger conversational responses. However, assertions are usually accompanied by a (non-conversational) Likert-type scale. Respondents are not accustomed to use such words, and therefore such response scales yield a high number of mismatch answers.

In the next chapter these factors will be elaborated into hypotheses that test the relation between question wording and the occurrence of mismatch answers. The hypotheses will be tested with non-experimental (chapter 6) and experimental survey data (chapter 7).

6 Non-experimental study: the relation between question characteristics and mismatch answers in existing data

6.1 Introduction

In chapter 5 we showed that mismatch answers from respondents, i.e., answers that are not clearly formatted as the scripted response alternatives, appear to be the most frequently occurring problematic deviations from the paradigmatic QA-sequence. Mismatch answers are also the most important cause of problematic deviations initiated by interviewers. As Smit (1995) concludes, interviewers often try to repair respondents' inadequate answers, and suggestive probing seems to be the most 'effective' way.

According to Houtkoop-Steenstra (2000), mismatch answers (or, as she labels them, 'unformatted' answers) are remarkable for two reasons. "From a cognitive perspective, one would expect that it is much easier to merely repeat a line that was just presented by the interviewer than it is to formulate a different answer". And from a social perspective, "one might expect respondents to be willing to please the interviewer and to make her task as easy as possible" (p. 183). Nevertheless, mismatch answers do occur frequently, and Houtkoop-Steenstra argues that respondents may have a good reason not to provide formatted answers.
In chapter 5 we concluded that the wording of questions and response alternatives are important determinants for the occurrence of mismatch answers. We also mentioned two specific factors, i.e., cognitive and conversational ones, which may account for the occurrence of mismatch answers. Interestingly, these factors refer to the same processes as Houtkoop-Steenstra's factors accounting for the occurrence of adequate answers.

6.1.1 Cognitive mismatch answers

When information to answer a question is not readily available in memory, respondents are faced with *state uncertainty*, and more thorough cognitive processing is required (Schaeffer and Thomson 1992). The problem relates to the second and third step of Tourangeau et al.'s model of cognitive processing (see section 2.3 of chapter 2). The second step of this model (retrieving relevant information from memory) can be troubled by state uncertainty when information is not readily available from memory. The third step (forming a judgment from the retrieved information) can be troubled by state uncertainty when a respondent has difficulty in deciding on the importance or relevance of the retrieved information in view of the question at hand.

Due to state uncertainty, respondents are not able to immediately produce an adequate answer, whereas they realize that the interviewer is waiting for an answer. Consequently respondents tend to give verbal considerations, i.e., a kind of thinking aloud. Dijkstra and Ongena (forthcoming) show that these verbal considerations are likely to be followed by an answer that usually is not adequate (a mismatch answer). The occurrence of verbal considerations and mismatch answers could be caused by question characteristics (some

questions are more difficult than others), but of course respondents' characteristics are involved as well (for example the older, less educated respondents, may have more difficulty with answering questions than others). Even interviewers' characteristics may be involved as well, for example the extent to which an interviewer is perceived as impatient.

Excerpt 6-1 gives an example of a mismatch answer that is likely to be caused by cognitive factors. The required answering format is a percentage. In line 2 the respondent comments on the difficulty of the task, in line 3 he gives verbal considerations and in line 4 he gives a cognitive mismatch (i.e., an answer formatted in hours instead of a percentage).

Excerpt 6-1 Q-A sequence with an example of a cognitive mismatch answer*

1	I: And uhm what percentage of time did you watch with attention?
2	R: That's hard
3	R: Well it is a quarter to ten so that would be five and a half hours (..) let's go from there
4	R: Well then it is three and a half hours
5	I: Uh that is a little bit more than fifty percent
6	R: Yes definitively
7	I: Sixty five percent?
8	R: Yes let's assume that
9	I: Yes

*This Q-A sequence was taken from the Television Survey that is described in chapter 5

6.1.2 Conversational mismatch answers

The second factor that may account for the occurrence of mismatch answers is a conversational problem. When information to answer a question is readily available (such as the respondent's birth date), it is still possible that mismatch answers occur, due to the fact that respondents may view and treat the interview as a kind of everyday conversation. In that case they are not necessarily motivated to give precisely formatted answers.

Although Schuman and Presser (1981) argue that most respondents "accept the framework of questions and try earnestly to work within that framework" (p. 299), the occurrence of conversational mismatch answers shows they do not always follow "the rules of the game".

Respondents are likely to give conversational mismatch answers if they are not focused on the task (i.e., giving precisely formatted answers). Again this may be caused by respondents' and question characteristics. For respondents' characteristics, age may again be relevant (older respondents might be less focused on the task than younger ones). For question characteristics, not the difficulty of questions, but the question wording may play a role. Dijkstra and Ongena (forthcoming) showed that especially assertions (with a Likert-type scale) received a lot of mismatch answers consisting of just 'yes' or 'no'. Such responses are perfectly normal in ordinary conversations. Assertions, such as "Commercials are funny to look at", are rather conventional expressions. Therefore, a conversational style of responding

(i.e., providing yes-no answers instead of agree-disagree answers) may be triggered by questions that resemble expressions that are common in conversations.

Excerpt 6-2 gives an example of a conversational mismatch. The required answering format is one of four categories as indicated by the interviewer. We can classify the respondent's utterance in line 2 as a mismatch answer. The problem is not so much caused by state uncertainty, because the respondent does not indicate problems in retrieving an answer. The respondent gives an answer that is perfectly normal in ordinary conversations, and also elaborates on his answer. Elaborations are not useful in survey interviews (unless they indicate problems in misunderstanding of questions that can subsequently be solved), but they may be regarded as cooperative in ordinary conversations (see section 2.2.15 of chapter 2). Thus, the problem seems to be caused by a lack of focus on the task of giving precisely formatted answers. Therefore, we classify the mismatch answer as a conversational mismatch.

Excerpt 6-2 Q-A sequence with an example of a conversational mismatch answer

1. I: Do you strongly agree, agree, disagree or strongly disagree with the assertion "For me, television commercials are too much alike"?
2. R:Well yes, they are all the same
3. R: Especially those detergent commercials, I get tired of them

Conversational mismatch answers are also related to the principle of satisficing (Krosnick 1999). In case of satisficing, respondents may select response alternatives without taking the effort to generate an optimal answer. This will result in adequately formatted, but possibly incorrect answers. Respondents may for the same reasons, as a conversational kind of satisficing, not even take the effort to select one of the response alternatives, but merely give a conversational mismatch answer. They may give only part of a response alternative (i.e., uttering only 'fairly' in case of response alternatives including 'fairly agree' and 'fairly disagree'). Apparently, respondents do not realize that the interviewer is not able to score the answer when it is not precisely formatted.

Conversational mismatch answers may especially occur for several questions in a row with the same response alternatives (e.g., batteries of assertions). For example, respondents can answer with phrases like 'the same'. With such a response it is not clear whether the respondent means 'the same as the previous question' or perhaps the same as questions earlier than the one immediately preceding. Furthermore, even when the question to which the respondent refers is adequately identified, it is possible that the questions are formulated oppositely. For example, the respondent agreed on the previous, negatively formulated question, and the next question is positively formulated. A 'same' answer on the latter question may mean the same alternative ('agree') or the same direction ('disagree'). The occurrence of a conversational mismatch may have a motivational cause: respondents are not willing to memorize the listed response alternatives, and this clearly refers to their lack of focus on the task of answering adequately.

6.1.3 Task mismatch answers

The example in Excerpt 6-3, taken from Schaeffer and Maynard (2002, p. 268), illustrates another situation that may occur when information is readily available, yet the respondent is not able to provide an adequate answer.

Excerpt 6-3 Q-A sequence with an example of a task mismatch answer

1.	I: Do you have your own business or farm?
2.	R: Weahh well I'm in partnership with my sister in the shoe business
3.	I: Okay so that would qualify as your own business?
4.	R: I guess so
5.	I: uh huh

In line 2 the respondent gives a report, which consists of pieces of relevant information. According to Schaeffer and Maynard (2002, p. 268) "the reports appear to accompany an uncertainty on the respondent's part –not about the facts that are detailed in the reports (which are produced with no signs of hesitation or markers of uncertainty)– but about how to translate those details into the categories of the survey question. That is, the problem seems to be a kind of task uncertainty, and not state uncertainty".

It is very likely that these reports are followed by mismatch answers (the utterance in line 4 may be classified as a qualified mismatch answer). Dijkstra and Ongena (forthcoming) found the same pattern, reports followed by mismatch answers, for cognitive mismatch answers. However, cognitive mismatch answers are related to *state* uncertainty, not *task* uncertainty. Thus, a third type of mismatch answers is distinguished: 'task' mismatch answers.

Task uncertainty is also related to the cognitive steps of answering a survey question of Tourangeau, et al.'s (2000) model. However, in this case the first and last step are involved. The first step (understanding the question) can be troubled by task uncertainty when ambiguous concepts are included in the question. The last step (formatting the response) can also be troubled by task uncertainty. When the information required by the question is available, but the respondent has difficulty to translate this into response alternatives (and in fact this might be a matter of questionnaire design), a problem of task uncertainty exists and task mismatch answers are likely to occur. Although the respondent is focused on the task of giving precisely formatted answers, a problem exists in translating the (probably complex) situation into the formatted response categories.

6.2 Hypotheses

6.2.1 Conversational and formal questions

Cannell et al.'s findings (1977) showed that respondents generally do not know what is expected of them, and argued that the questionnaire and the techniques interviewers use should clarify the respondents' general task; provide cues as to how respondents can answer questions most efficiently, and motivate them to meet requirements of accurate responses. The specific instructions they used in their experiment yielded more precise and elaborate reports. These instructions clarified the goal of the survey and provided cues that clarified intended task performance (e.g., "we'd like you to be as exact as you can").

However, wording of the question proper may also be important with respect to hinting at the respondent's task. A question may signal respondents about the character of the survey. We assume that conversationally worded questions, in which the wording of ordinary conversations is used, may give false signals to the respondent about the required degree of accuracy in reporting, whereas formally formulated questions alert the respondent to the formal character of the survey, reminding them to answer with precisely formatted answers.

Thus, the first hypothesis is:

> *H1 A question that is formulated as a conversational question will generate more mismatch answers, than a formally worded question*

6.2.2 Conversational and formal alternatives

In addition, conversational mismatch answers are less likely to occur when the response alternatives are worded as conversational responses. For respondents it is much easier to produce an answer that is a normal expression in conversations. For example, it is much easier for respondents to answer a question with 'yes' or 'no', than to use the words 'agree' or 'disagree'. A probable cause for the trouble with uncommon formal alternatives may be a disruption in cognitive processing. Holbrook et al. (2000) found that unconventional response orders also caused a disruption in cognitive processing. The disruption caused by unconventional response orders or (in our case) unconventional words may (temporarily) distract respondents from their response task, and as a consequence they may loose their focus on the exact wording of the alternatives. In addition, when alternatives are used for several questions in a row (as is often the case for batteries of assertions), alternatives are often presented only at the introduction of the battery. Thus, the alternatives have to be stored in short-term memory. It is likely that this storage is easier for frequently used (i.e., conversational) words than it is for uncommon (i.e., formal) words.

Thus, the second hypothesis is:

H2a Questions with conversational response alternatives will generate less mismatch answers, than the same questions with comparable formal response alternatives.

6.2.3 Implicit and listed alternatives

A question that asks for a number (for example a number of hours, minutes, days, months, years etc.) is one of a specific type. We consider such a question a closed question, with implicit alternatives. Such a question in fact implies a range of response alternatives. For example, for a question like 'How many days a week do you watch television?' the response alternatives are limited from zero to seven days. The response alternatives are implicit because the question gives an indication of what kind of response alternative is required, but does not explicitly list them.

Questions posed in ordinary conversations can often be regarded as questions with implicit alternatives. For example, 'What is your age?' implies that the answer must be a number of years. 'How long will this journey by train take?' implies that the answer must be a number of hours and/or minutes or days. Actually, in ordinary conversations, yes-no questions and open questions are commonly asked, and it would be awkward to ask questions with a set of alternatives (e.g., 'Does this journey by train take less than twenty minutes, between twenty and forty minutes or more than forty minutes?'). The format of questions with implicit alternatives can thus be considered more conversational than listed alternatives.

It is very important that the alternatives in the questionnaire correctly match the alternatives implied by the question. When for questions with implicit alternatives the actual alternatives do not match the implied alternatives, these questions are improper questions. Such questions imply other alternatives than the alternatives that are actually used. For example, a question like 'how often do you watch television?' implies alternatives such as 'very often', 'not so often' and 'hardly ever'.

Consider the following questions:

(a) 'What is the number of days in a week that you watch television? ….days.'
(b) 'Do you, on zero, one, two, three, four, five, six or seven days a week, watch television?'
(c) 'Do you, every day, most days, some days or hardly ever watch television?'

With questions (b) and (c) [12] the respondent is explicitly informed about the response alternatives, but we do not know to what extent the respondent is able to grasp the idea of

[12] The questions (b) and (c) have, in English, an uncommon grammatical structure. A more logical structure would have been to start the questions with 'Do you watch television…'. However, in that case the question

using these alternatives to answer the question. The problem is that, on the one hand the question delivery component (i.e., 'do you watch television') should be presented last, in order to prevent the respondent from interrupting the interviewer before she has read the alternatives. On the other hand, the question delivery component should not be presented last, because then the respondents do not know what the question is going to be about, and therefore have trouble to pay attention to the long list of alternatives. The long list that seems to go nowhere may create confusion, which distracts respondents from the idea that they have to use the alternatives in order to answer the question.

With question (a), the respondent is not explicitly informed about the response alternatives, but this question very obviously implies that a number is required for an answer. Therefore, question (a) is likely to yield less mismatch answers than a question like (b), but also less mismatch answers than a question with conversational alternatives such as (c).

We thus add to hypothesis 2a the following hypothesis:

H2b Questions with listed (i.e., conversational or formal) response alternatives will generate more mismatch answers than questions with implicit response alternatives

6.2.4 Difficult and easy questions

In case of questions requiring substantive cognitive processing, respondents may be well aware of the inadequacy of their cognitive mismatch answer. Respondents may not have enough information to decide between response alternatives. This lack of information is caused by a problem of retrieval of adequate information from their memory. Such 'deep' cognitive processing often results in spontaneous verbal considerations uttered during answering the question at hand. While respondents are giving these considerations, they are liable to become distracted from the response task of giving precisely formatted answers. They are also apt to end up giving some estimation referring to multiple response categories.

Thus, our third hypothesis is:

H3 Questions requiring information not readily available in memory (i.e., difficult questions) will generate more mismatch answers, than questions requiring relatively little cognitive processing (i.e., easy questions).

6.2.5 Ambiguous and non-ambiguous questions

Task mismatch answers occur when the respondent is faced with some ambiguity in the response task. They may, like cognitive mismatch answers, be caused by a lack of information to decide between alternatives. However, this lack of information is assumed to

---

component (i.e.,'do you watch television') is presented, before the response alternatives are read, which is then likely to be interrupted. The Dutch equivalents of these questions are grammatically correct.

be caused by an ambiguity captured in the question and respondents' own information regarding the question, and not because of a lack of adequate information from memory. Task mismatch answers can therefore be avoided by unambiguous question wording. For example, the task mismatch answer, after the question 'Do you have your own business or farm?' in Excerpt 6-3 could have been avoided when the question would read 'Do you, either alone or with others, have your own business or farm?'

Thus, our fourth hypothesis is:

H4 Questions containing ambiguous concepts will generate more mismatch answers than questions not containing ambiguous concepts.

6.3 Non-experimental study

A non-experimental study was conducted to investigate whether the different types of questions (conversational, formal, easy, difficult, ambiguous and non-ambiguous questions) and alternatives (conversational, formal, implicit and listed) could be identified in a normal survey, and were related to the occurrence of mismatch answers. We also aimed to investigate whether the three types of mismatch answers (conversational, task and cognitive mismatch answers) could be identified, and could also be related to question wording and types of alternatives. For these analyses Q-A sequences from the Dutch pilot of the European Social Survey (ESS) were used. In order to test our hypotheses, a large number of different questions that comprise examples of all question types, is very useful. One important advantage of the ESS data was that the questionnaire fulfilled this criterion.

6.3.1 The data

The interviews of the Dutch pilot study of the ESS concern face-to-face interviews that were conducted, by means of a CAPI program, in the spring of 2002 (ESS 2005). The questionnaire consisted of 268 questions. Tapes were available from seven interviewers. The taped CAPI interviews were digitized. It turned out that 23 interviews with a good recording quality were available. Due to time constraints, not all 268 questions of all interviews were transcribed. Eight interviews were transcribed completely, for the other 15 interviews at least the first 100 questions of the interview were transcribed. Three questions (A2, H2 and H4) were hardly ever asked (i.e., the gender, country of birth and nationality of the respondent), and for that reason not included in the analyses.

Coding of the data was done by three different coders. To test the reliability of the coding, we randomly selected 1100 Q-A sequences that were initially coded by two of the coders, to be coded again by the third coder. Comparison of the twice-coded 1100 Q-A sequences yielded a Kappa of 0.70. The dataset eventually coded and used for analysis was checked for unlikely and rare codes, which may have resulted in a higher validity of the coding.

In Table 6-1 the numbers of Q-A sequences that were coded, or excluded for various reasons, are shown. The total number of in- and excluded Q-A sequences (6164) equals the 268 questions multiplied by the 23 respondents. It turns out that the number of incorrectly skipped questions is quite high, considering the fact that the survey was administered on a computer. This was particularly due to the behavior of one interviewer.

Table 6-1 Number and percentage of Q-A sequences coded in the ESS-data

	Number of Q-A sequences	Percentage
Asked and coded	3623	58.8%
Correctly skipped	353	5.7%
Incorrectly skipped	250	4.1%
Part of interview not on tape	87	1.4%
Not coded	1851	30.0%
Total	6164	100.0%

6.3.2 Different types of question wording

In order to non-experimentally test the hypotheses as described in section 6.2, we distinguished versions of each question type among the 268 ESS questions: conversational and formal questions, easy (requiring less cognitive effort) and difficult questions (requiring more cognitive effort) and ambiguous and non-ambiguous questions. Furthermore, three types of alternatives were distinguished: conversational, formal and implicit response alternatives (i.e., not explicitly listed). Open questions (n=7), and the three questions that were hardly ever asked (see section 6.3.1) were not considered, thus 258 questions were included.

The distinctions resulted in 3*2*2*2=24 possible combinations, i.e.,: three types of alternatives for hypotheses 2a and 2b and two types of questions for each of the three hypotheses 1, 3, and 4). In the ESS questionnaire, 16 different question types could actually be distinguished instead of 24. Table 6-2 shows in a 2*3 form some examples of questions categorized for their conversational character and types of alternatives. For formal questions (75% of the questions) all three types of alternatives (formal, conversational and implicit alternatives) were available. For conversational questions (25% of the questions) the alternatives were all categorized as formal. Hence, we could not distinguish between formal and conversational alternatives for conversationally worded questions.

Table 6-2 Examples of formal and conversational questions and alternatives in the ESS questionnaire

	Formal question	*Conversational question*
Formal alternatives	*(n = 111 questions)* To what extent do you consider yourself associated with this party? 1 Very associated 2 Fairly associated 3 Hardly associated 4 Not at all associated I think I can play an active role in a group that is focused on political issues 1 Strongly agree 2 Agree 3 Neither agree nor disagree 4 Disagree 5 Strongly disagree	*(n = 67 questions)* Taking all things together, are you 1 Very happy 2 Fairly happy 3 Not so happy 4 Not at all happy Politicians do not care what people like me think 1 Strongly agree 2 Agree 3 Neither agree nor disagree 4 Disagree 5 Strongly disagree
Conversational alternatives	*(n = 61 questions)* Do you consider yourself as a member of a minority group that is discriminated in this country? 1 Yes 2 No Compared to other people of your age, how often do you take part in social activities? 1 Much less than most 2 Less than most 3 About the same 5 More than most 6 Much more than most	_ _
Implicit alternatives	*(n = 19 questions)* In which year were you born?	_

The distinctions between formal and conversational questions were made on the basis of the conversational character of the question. When question wording was considered to include common words and a sentence structure that is generally normal to use in ordinary conversations, the question was categorized as conversational. All other questions were considered formal.

An assertion such as 'Politicians do not care what people like me think' was considered to consist of common words and to be a normal expression in ordinary conversations. A question like 'Do you consider yourself as a member of a minority group that is discriminated in this country?' is not likely to be formulated as such in ordinary conversations, and is categorized as formal. Of course, there are questions that can be considered far more

conversational than other questions within the same category, but we chose to dichotomize the questions into the two categories that the hypotheses refer to, and not to complicate the categorization with gradations of the conversational character of the questions.

In the same way we categorized the types of alternatives; alternatives that consisted of common words and a simple structure (e.g., 'yes' and 'no') were considered conversational alternatives.

Table 6-3 shows in a 2*2 form, examples of questions that were classified as non-ambiguous or as ambiguous, and as difficult or as easy.

Table 6-3 Examples of ambiguous, non-ambiguous, easy and difficult questions in the ESS questionnaire

	Ambiguous	*Non-ambiguous*
Easy	*(n = 91 questions)* Including yourself, how many people live here regularly as member of this household?	*(n = 112 questions)* In which year were you born?
Difficult	*(n = 21 questions)* On an average weekday, how much time do you generally spend watching television?	*(n = 34 questions)* Out of every 100 people living in the Netherlands, how many do you think were born outside the Netherlands?

Questions were assessed as ambiguous when they included concepts that were not specified. Non-ambiguous questions do not include ambiguous concepts, or they are specified. For example, the concept 'regularly' in the question 'how many people live here regularly as a member of this household' is not specified, and accordingly the question is considered ambiguous. The difficulty of questions was assessed by determining the relative amount of cognitive processing required to answer the question. We were primarily concerned with steps 2 and 3 of Tourangeau et al.'s model, i.e., information retrieval and judgment. For example, most behavioral frequency questions were considered difficult, since they required respondents to retrieve information about the number of hours and minutes spent on some behavior for an entire week.

6.3.3 Mismatch answers and conversational character of questions (H1 and H2)

Table 6-4 shows the frequency of mismatch answers that occur for each of the four possible question types, concerning the conversational character of question wording and the three types of alternatives.

Table 6-4 Frequency of mismatch answers for four question types

	Conversational Q		Formal Q					
	Formal alternatives		Formal alternatives		Conversational alternatives		Implicit alternatives	
Q-A sequences								
With mismatch	114	11%	251	16%	46	6%	46	18%
Without mismatch	926	89%	1326	84%	708	94%	206	82%
Total	1040	100%	1577	100%	754	100%	252	100

Taking only formal alternatives into account, conversational questions (11%) do not yield more mismatch answers than formal questions (16%). Hypothesis 1 can thus not be confirmed. However, hypothesis 2a can be confirmed: formal questions with conversational alternatives yield less mismatch answers (6%) than formal questions with formal alternatives (16%, $\chi^2 = 44.20$, df = 1, $p < 0.01$).

Hypothesis 2b cannot be confirmed; implicit alternatives yield most (18%) instead of least mismatch answers. This may be explained by the fact that questions with implicit alternatives mostly (i.e., 67%) concerned rather difficult questions. Hence, the high percentage of mismatch answers is likely to be due to cognitive mismatch answers. It may be possible that the use of implicit alternatives has nevertheless reduced the chance of conversational mismatch answers, but that this reduction is compensated by a higher chance of cognitive mismatch answers.

6.3.4 Confounding question characteristics (H1 and H2)

The analyses in Table 6-4 concerned all questions, i.e., those with and those without a show card. The use of show cards can decrease the occurrence of mismatch answers (e.g., Dijkstra and Ongena forthcoming; Prüfer and Rexroth 1985). Therefore we have to examine a possible confounding of the question types and the use of show cards. As is shown in Table 6-5, show card questions were not equally distributed over the types of questions and alternatives.

Table 6-5 Percentage of questions with a show card for types of questions and alternatives

	Conversational Q		Formal Q					
	Formal alternatives		Formal alternatives		Conversational alternatives		Implicit alternatives	
Questions								
Without show card	8	12%	28	25%	47	77%	15	79%
With show card	59	88%	83	75%	14	33%	4	21%
Total	67	100%	111	100%	61	100%	19	100%

From the unequal distribution of show cards over the different question types we can conclude that it is dubious to compare all questions without controlling for show cards. For example, questions that were categorized as conversational questions with formal alternatives are less often (12%) accompanied by a show card than questions that were categorized as

formal questions with formal alternatives (i.e., 25%). Thus, including show card questions to test the effects of type of questions (hypothesis 1) is doubtful. The unequal distribution of show cards for questions with conversational alternatives, formal alternatives, and implicit alternatives, also makes testing hypotheses 2a and 2b dubious.

We could choose to perform separate analyses for questions with and questions without show cards. However, the effects of question wording for show card questions are difficult to interpret. A show card is likely to alert the respondents to the formal character of the survey, and the most useful tool to remind them to answer with precisely formatted answers. When respondents nevertheless give mismatch answers, this may be due to an ambiguous character of the question or response alternatives, due to the difficulty of the question or just because the respondent has received the wrong show card. Therefore we will test the hypotheses again for questions without show cards only. As is shown in Table 6-6, the results for hypothesis 1 and 2 are different when questions with show cards are excluded as compared to the results for all questions (in Table 6-5).

Table 6-6 Frequency of mismatch answers occurring for different types of questions (considering only questions without show cards)

	Conversational Q		Formal Q					
	Formal alternatives		Formal alternatives		Conversational alternatives		Implicit alternatives	
Q-A sequences								
With mismatch	32	26%	63	17%	15	2%	41	20%
Without mismatch	90	74%	299	83%	556	97%	165	80%
Total	122	100%	362	100%	571	100%	206	100

For questions without show cards and with formal alternatives only, conversational questions yield more mismatch answers (26%) than formal questions (17%, $\chi^2 = 3.04$, df = 1, $p < 0.05$). Questions with formal alternatives yield more mismatch answers (16%) than questions with conversational alternatives (2%, $\chi^2 = 63.28$, $p < 0.01$). These results confirm hypothesis 1 and 2a respectively. Hypothesis 2b could not be confirmed: questions with implicit alternatives still yield a higher percentage of mismatch answers than questions with formal or conversational alternatives.

We have to keep in mind that the questions were not designed in a proper split ballot experiment. The questions do not only differ with respect to characteristics such as show cards, formal and conversational wording, but also with respect to the topics being asked about, and the order of occurrence in the questionnaire. For example, all background questions were categorized as formal questions (e.g., respondents' age, number of persons in the respondents' household, etc.). It may have been possible that the topic of the questions has influenced the percentage of mismatch answers.

A distinction that is indirectly related to the content of questions is the one between perception, factual and opinion questions. A perception question concerns the respondent's judgment of his own state, concerning concepts such as health, happiness, etc. An example of a perception question is 'Taking all things together, are you, very happy, fairly happy, not so

happy, or not at all happy?' An example of an opinion question is an assertion like 'Politicians do not care what people like me think'. An example of a factual question is 'In which year were you born?' These three types of questions were very unequally distributed over formal and conversational questions and alternatives, as is shown in Table 6-7. Some question type-combinations yield empty cell values, which makes testing hypotheses controlling for perceptual, factual and opinion questions difficult or impossible.

Table 6-7 Percentage of perceptual, factual and opinion questions for conversational and formal questions and types of alternatives

	Conversational Q		Formal Q					
	Formal alternatives		Formal alternatives		Conversational alternatives		Implicit alternatives	
Questions								
Perception	3	38%	0	0%	3	6%	1	6%
Factual	4	50%	23	82%	44	94%	14	93%
Opinions	1	13%	5	18%	0	0%	0	0%
Total	8	100%	28	100%	47	100%	15	100%

In Table 6-8 the percentage of mismatch answers is shown for the question types within perception, factual and opinion questions.

For opinion questions with formal alternatives (section C), the difference in the percentage of mismatch answers between the conversational question and the formal questions (n = 5 questions) is according to our expectation, However, the number of cases is too small to test for statistical significance.

It appears that for factual questions with formal alternatives, conversational questions do not yield more mismatch answers than formal questions. Within perception questions (section A) we cannot correctly test hypothesis 1, as conversational questions can only be compared to formal questions with the same type of alternatives, but none of the formal questions with formal alternatives concerned perception questions. Furthermore, the number of cases often is too small to be able to perform statistical tests.

Furthermore, testing hypothesis 2b, for factual questions, there is a significant difference in the percentage of mismatch answers between questions with conversational and implicit alternatives (i.e., 2% versus 21 %, $\chi^2 = 80.92$, df = 1, $p < 0.01$). This difference is contrary to our hypothesis.

These results show that the effects of question wording are very likely to depend on the type (i.e., perception, factual or opinion) of question.

Table 6-8 Percentage of mismatch answers for perceptual, factual and opinion questions

A. Perception Q	Conversational Q		Formal Q					
	Formal alternatives		Formal alternatives		Conversational alternatives		Implicit alternatives	
Q-A sequences								
With mismatch	17	30%	-	-	7	11%	1	7%
Without mismatch	39	70%	-	-	56	89%	13	93%
Total	56	100%	-	-	63	100%	14	100%
B. Factual Q	Conversational Q		Formal Q					
	Formal alternatives		Formal alternatives		Conversational alternatives		Implicit alternatives	
Q-A sequences								
With mismatch	7	13%	43	17%	8	2%	40	21%
Without	45	87%	212	83%	500	98%	152	79%
Total	52	100%	255	100%	508	100%	192	100%
C. Opinion Q	Conversational Q		Formal Q					
Q-A sequences	Formal alternatives		Formal alternatives					
With mismatch	5	36%	20	19%				
Without mismatch	9	64%	87	81%				
Total	14	100%	107	100%				

6.3.5 Mismatch answers and difficulty of questions (H3)

Table 6-9 shows, for questions without show cards, that mismatch answers occur relatively more often in Q-A sequences concerning difficult questions, than those concerning easy questions ($\chi^2 = 12.70$, $p < 0.01$). The results show support for hypothesis 3.

Table 6-9 Percentage of mismatch answers in difficult versus easy questions

Q-A sequences	Difficult Q		Easy Q	
With mismatch	58	17%	90	10%
Without mismatch	279	83%	834	90%
Total	337	100%	924	100%

$\chi^2 = 13.03$, df = 1, $p < 0.01$

In section 6.3.3 it was already mentioned that questions with implicit alternatives (e.g., the number of years respondents have lived in their neighborhood, or the number of hours or minutes they watch television on an average day) were primarily difficult questions. As is shown in Table 6-10, questions with implicit alternatives were categorized as difficult more often, than other types of questions. Thus, it may be useful to test hypothesis 3 separately for the different types of questions and alternatives and to test 1, 2a and 2b separately for easy and difficult questions.

Table 6-10 Percentage of difficult and easy questions for conversational and formal questions and types of alternatives

	Conversational Q		Formal Q					
Questions	Formal alternatives		Formal alternatives		Conversational alternatives		Implicit alternatives	
Difficult questions	1	13%	6	21%	8	17%	10	67%
Easy questions	7	87%	22	79%	39	83%	5	33%
Total	8	100%	28	100%	47	100%	15	100%

In Table 6-11 the percentage of Q-A sequences with mismatch answers for conversational and formal questions and the types of alternatives is shown in separate sections for difficult and easy questions. We first test hypothesis 3 again, while keeping the conversational character of the questions constant. This entails comparison across the sections A (difficult questions) and B (easy questions) of this table.

Taking only conversational questions with formal alternatives into account, the difficult question yielded more mismatch answers (43%) than the easy questions (21%). This difference is statistically significant in a one-tailed test ($\chi^2 = 3.18$, df = 1, $p < 0.05$). However, in case of formal questions (and formal alternatives), difficult questions yield less mismatch answers (9%) than easy questions (20%, ($\chi^2 = 5.35$, df = 1, $p < 0.05$). Difficult questions with conversational alternatives yield more mismatch answers than easy questions.

Finally, difficult questions with implicit alternatives not only occur more frequently, they also yield more mismatch answers (25%) than easy questions with implicit alternatives (6%, $\chi^2 = 9.08$, df = 1, $p < 0.01$). Thus, the effect of question difficulty is not entirely due to difference in the frequency of implicit alternatives for easy and difficult questions, and except for formal questions, we can confirm hypothesis 3 for all question types.

Table 6-11 Percentage of mismatch answers for different types of alternatives of difficult and easy questions

A. Difficult questions	Conversational Q		Formal Q					
Q-A sequences	Formal alternatives		Formal alternatives		Conversational alternatives		Implicit alternatives	
With mismatch	6	43%	7	9%	7	8%	38	25%
Without mismatch	8	57%	73	91%	83	92%	115	75%
Total	14	100%	80	100%	90	100%	153	100%
B. Easy questions	**Conversational Q**		**Formal Q**					
Q-A sequences	Formal alternatives		Formal alternatives		Conversational alternatives		Implicit alternatives	
With mismatch	23	21%	56	20%	8	2%	3	6%
Without mismatch	85	78%	226	80%	473	98%	50	94%
Total	108	100%	282	100%	481	100%	53	100%

Next, we test hypotheses 1, 2a and 2b again within the difficult and easy questions. In case of difficult questions with formal alternatives, conversational questions yield more mismatch answers (43%) than formal questions (9%, $\chi^2 = 11.63$, df = 1, $p < 0.01$), thus hypothesis 1can be confirmed for difficult questions. There is no difference in the percentage of mismatch answers for formal versus conversational alternatives, but implicit alternatives yield more mismatch answers than formal alternatives ($\chi^2 = 8.73$, df = 1, $p < 0.05$), and also more mismatch answers than conversational alternatives ($\chi^2 = 10.93$ df = 1, $p < 0.01$). Thus, hypothesis 2a cannot be confirmed, and we found results contrary to hypothesis 2b for difficult questions.

For easy questions with formal alternatives, there is no difference in the percentage of mismatch answers for conversational and formal questions. However, considering formal questions, formal alternatives yield more mismatch answers than conversational alternatives ($\chi^2 = 76.6$, df = 1, $p < 0.01$, i.e., confirming hypothesis 2a)

Furthermore, easy questions with implicit alternatives do not yield less mismatch answers than easy questions with conversational alternatives, but implicit alternatives yield less mismatch answers (6%) than formal alternatives (20%, $\chi^2 = 6.2$, df = 1, $p < 0.05$). One of those easy questions with implicit alternatives concerned the respondent's year of birth. This question yielded no mismatch answers at all. The question of course concerns information that is likely to be readily available for respondents. It also is a clear example of a question that precisely implies the required response format. Thus, only for easy questions that are accompanied by the formal versus implicit alternatives, hypothesis 2b can be confirmed.

In summary, these differences show that the effects of the conversational character of questions are different for difficult and easy questions. For difficult questions we could confirm hypothesis 1, that conversational questions generate more mismatch answers than formal questions, but not for easy questions. For easy questions we could confirm hypothesis 2a, that formal alternatives generate more mismatch answers than conversational alternatives, but not for difficult questions. Finally, for difficult questions we found results opposite to hypothesis 2b, that listed alternatives generate more mismatch answers than implicit alternatives, and for easy questions we could only confirm that questions with implicit alternatives yield less mismatch answers than questions with formal alternatives.

6.3.6 Mismatch answers and ambiguity of questions (H4)

Table 6-12 shows that ambiguous questions yield more mismatch answers than non-ambiguous questions ($\chi^2 = 3.80$, $p < 0.05$). Although the difference in the percentage of mismatch answers is fairly small, this result confirms hypothesis 4.

Table 6-12 Percentage of mismatch answers in non- ambiguous versus ambiguous questions

Q-A sequences	Non-ambiguous Q		Ambiguous Q	
With mismatch	72	10%	76	14%
Without mismatch	640	90%	473	86%
Total	712	100%	549	100%

$\chi^2 = 4.17$ df = 1, $p < 0.05$

As is shown in Table 6-13 the ambiguity of questions is related to the conversational character of questions and alternatives, but less strongly than was the case for difficulty of questions. Formal questions (with formal alternatives and conversational alternatives) were more often categorized as non-ambiguous than conversational questions. This may indicate an overlap in the operationalization of ambiguity and the conversational character of questions: in ordinary conversations concepts are often not specified, thus when ambiguous concepts in questions are specified in questions, they are more likely to be categorized as formal.

Table 6-13 Percentage of non-ambiguous and ambiguous questions for conversational and formal questions and types of alternatives

	Conversational Q		Formal Q					
	Formal alternatives		Formal alternatives		Conversational alternatives		Implicit alternatives	
Questions								
Non-ambiguous	4	50%	21	75%	29	62%	8	53%
Ambiguous	4	50%	7	25%	18	38%	7	47%
Total	8	100%	28	100%	47	100%	15	100%

Table 6-14 shows in separate sections for non-ambiguous and ambiguous questions the percentage of Q-A sequences with mismatch answers for conversational and formal questions and the types of alternatives. When we test hypothesis 4 again, while holding the conversational character of questions constant, i.e., with a comparison of the sections A (non-ambiguous questions) and B (ambiguous questions), this yields no significant differences. Thus, we cannot confirm hypothesis 4 within any of the question types.

When we test hypothesis 1 again, while holding the ambiguous character of questions constant, i.e., comparisons within sections A an B, it appears that the effects of conversational character of questions do not differ for non-ambiguous questions versus ambiguous questions. For both non-ambiguous and ambiguous questions, there is, within the formal alternatives, no significant difference in the percentage of mismatch answers between conversational and formal questions. Questions with implicit alternatives yield more mismatch answers than conversational alternatives ($\chi^2 = 30.04$, df = 1, $p < 0.01$ and $\chi^2 = 27.81$, df = 1, $p < 0.01$ for sections A an B respectively). Questions with formal alternatives

yield more mismatch answers than conversational alternatives ($\chi^2 = 47.68$, df = 1, $p < 0.01$ and $\chi^2 = 18.57$, df = 1, $p < 0.01$ for sections A and B respectively).

Table 6-14 Percentage of mismatch answers for different types of alternatives of non-ambiguous and ambiguous questions

A. Non-ambiguous Q	Conversational Q		Formal Q					
	Formal		Formal		Conversational		Implicit	
Q-A sequences	alternatives		alternatives		alternatives		alternatives	
With mismatch	19	27%	39	17%	4	1%	10	14%
Without mismatch	51	73%	195	82%	334	99%	60	86%
Total	70	100%	234	100%	338	100%	70	100%
B. Ambiguous Q	Conversational Q		Formal Q					
	Formal		Formal		Conversational		Implicit	
Q-A sequences	alternatives		alternatives		alternatives		alternatives	
With mismatch	10	19%	24	19%	11	5%	31	23%
Without mismatch	42	81%	104	81%	222	95%	105	77%
Total	52	100%	128	100%	233	100%	136	100%

To summarize, hypothesis 1, 2a, 3 and 4 could generally be confirmed for the questions without show cards, but differences existed when controlled for other aspects of questions. The independent variables were correlated, which complicated the interpretation of results. For example, hypothesis 1 could not be confirmed for easy questions, hypothesis 2a could not be confirmed for difficult questions, hypothesis 3 could not be confirmed for formal questions and hypothesis 4 could not be confirmed when controlled for the conversational character of questions and the types of alternatives. Hypothesis 2b (concerning implicit versus listed alternatives) could not be confirmed at all, and even yielded contrary results for difficult questions.

The hypotheses tested were focused on differences in the occurrence of specific types of mismatch answers. For example, hypothesis 2b was intended to test whether chances of mismatch answers were different for implicit or listed alternatives, based on a conversational cause of mismatch answers. The questions within the category of questions with implicit alternatives were more often categorized as difficult than the other question types. It may have been possible that the hypothesis was not confirmed because the chance of cognitive mismatch answers was higher for the questions with implicit alternatives than for the other questions. Taking these considerations into account, it would be useful to relate the type of mismatch answer to the effects of questions and alternatives. A distinction of the three types of mismatch answers is described in the next section.

6.4 Occurrence and recognition of three types of mismatch answers

In appendix 6-1 a summary is given of the distinction of the three types of mismatch answers in the ESS data. The way the three types of mismatch answers are recognized seems to be rather subjective; from information that respondents provide in their answer or in addition to their answer, the type of mismatch answer was inferred.

Conversational mismatch answers are mainly recognized by answers that (1) are produced without hesitation. Respondents are convinced about the idea that their answer is acceptable. They are having a conversation and do what they normally do, without being evaluated for the preciseness of their responses. Furthermore, when they respond in a conversational style, respondents may (2) provide elaborations, usually after their answer (as Excerpt 6-2 in section 6.1 illustrated).

Task mismatch answers are recognized by (1) hesitations and (2) qualifying statements. Respondents may (3) provide relevant information that indicates they are certain about the information they have retrieved, but not certain how this information can be mapped on to the response categories. The answer can also be accompanied by (4) a condition (e.g., with words like 'it depends'). Furthermore, a task mismatch answer can be recognized by (5) requests for clarification that precede or follow the answer.

Cognitive mismatch answers are indicated by (1) preceding think aloud utterances, and (2) filled pauses ('uhs'). Respondents may also indicate uncertainty, but for these mismatch answers this indicates (3) uncertainty about the adequacy of the retrieved information. The mismatch answer given may (4) comprise a range ('between 1 and 2 hours'), which indicates that respondents have used an estimation strategy to come up with an answer.

Quite often there is hardly any information in a Q-A sequence that indicates the cause of a mismatch answer. Especially when the information required by the question is likely to be immediately accessible, no information on the cause of the mismatch answer may be derived from the interaction. In those cases, a mismatch answer is considered a conversational mismatch. It is only possible to label a mismatch answer as task or cognitive mismatch when a task or cognitive problem is indicated by the factors mentioned above.

Table 6-15 shows the frequency of occurrence of the three types of mismatch answers. Questions with show cards yield less cognitive and task mismatch answers and more conversational mismatch answers than the questions without show cards. This result is difficult to interpret; we would expect show cards to remind the respondents of the formal character of the survey. As we have indicated before, it may have been possible that mismatch answers occurred because respondents were using the wrong show card.

Table 6-15 Frequency of occurrence of three types of mismatch answers.

	Questions without show cards		Questions with show cards		Total	
Type of mismatch answer	Freq.	%	Freq.	%	Freq.	%
Conversational	105	71%	274	89%	379	83%
Task	23	16%	33	10%	56	13%
Cognitive	30	14%	2	1%	22	5%
Total	148	100%	309	100%	457	100%

$\chi^2 = 40.13$ df = 2, $p < 0.01$

Table 6-16 shows the percentage of the three types of mismatch answers for the conversational and formal questions, and the three types of alternatives. The number of Q-A sequences involved is often too low to warrant statistical tests of significance. Therefore we present below the section for questions without show cards (A) and a section for all questions (B). It appears that our question distinctions are indeed related to the different types of mismatch answers. Although for all question distinctions the majority of the mismatch answers were recognized as a conversational mismatch answer, this percentage differs for the question types.

Table 6-16 Frequency of occurrence of three types of mismatch answers for question types

A. Questions without show cards	Conversational Q		Formal Q					
	Formal alternatives		Formal alternatives		Conversational alternatives		Implicit alternatives	
Type of mismatch								
Conversational	27	93%	47	75%	10	67%	21	51%
Task	2	7%	10	15%	5	33%	6	15%
Cognitive	0	0%	6	10%	0	0%	4	34%
Total	29	100%	63	100%	15	100%	41	100%
B. All questions	Conversational Q		Formal Q					
	Formal alternatives		Formal alternatives		Conversational alternatives		Implicit alternatives	
Type of mismatch								
Conversational	105	92%	212	84%	36	78%	26	57%
Task	8	7%	32	13%	10	22%	6	13%
Cognitive	1	1%	7	3%	0	0%	14	30%
Total	114	100%	251	100%	46	100%	46	100%

The mismatch answers that occur with conversational questions with formal alternatives are mostly conversational mismatch answers (93% for questions without show cards). Some of these questions comprised improper yes-no questions; the question was worded as a yes-no question, but the response alternatives comprised alternatives other than ‘yes’ and ‘no’ (see

section 6.5.3). For other conversational questions, a Likert-type scale of five response alternatives is often used, whereas respondents typically reply with 'yes' and 'no' as an answer.

Especially for formal questions with conversational alternatives, a relatively large part of the mismatch answers is categorized as a task mismatch answer (33%). Most of those questions with conversational alternatives concern yes-no questions, i.e., with only two alternatives. Such questions force respondents to choose between two extremes (like the 'Do you have your own business or farm'-question illustrated in Excerpt 6-3). It may be difficult for respondents to decide between two alternatives when they are not certain what situation applies to them. They may express this difficulty by means of a task mismatch.

The mismatch answers that occur after formal questions with implicit alternatives were relatively more often categorized as cognitive mismatch answers as compared to other question types. As we indicated in section 6.3.3, most of these questions concerned difficult questions. Questions with implicit alternatives were expected to have a lower chance of conversational mismatch answers. The cognitive processing required to answer these questions is likely to be expressed by means of considerations. These considerations increase the chance of (cognitive) mismatch answers.

As is shown in the first columns of Table 6-17, we also compared the distributions of the three types of mismatch answers for difficult versus easy questions. Difficult questions not only in general yielded more mismatch answers, but also more cognitive mismatch answers than easy questions.

For non-ambiguous questions versus ambiguous questions, we would expect that task mismatch answers (caused by ambiguity of question wording) would account for the difference found in the overall percentage of mismatch answers. However, as Table 6-17 shows, task mismatch answers did not occur more often with ambiguous questions than with non-ambiguous questions. This may be explained by the fact that task mismatch answers especially concerned problems that respondents had with the response alternatives, and not with question meaning. For example, respondents may find that the response alternatives are not detailed enough (see appendix 6-1).

Table 6-17 Percentage of occurrence of three types of mismatch answers for difficulty and ambiguity of questions

Type of mismatch	Difficult Q	Easy Q	Non-ambiguous Q	Ambiguous Q
Conversational	66%	88%	85%	79%
Task	14%	12%	13%	13%
Cognitive	20%	1%	3%	9%
Total	100%	100%	100%	100%

Overall, mismatch answers occurred in only 13% of 3623 Q-A sequences. The mismatch answers were the most frequently occurring problematic deviation as they accounted for 84% of the Q-A sequences with a problematic deviation. However, the overall percentage of

mismatch answers is low compared to surveys in general. For example, in the five surveys analyzed by Dijkstra and Ongena (forthcoming) the percentage of mismatch answers ranged between 12.1% and 31.4%. The relatively low percentage in the ESS can largely be explained by the fact that in this survey show cards were used. Show cards are a useful tool to decrease the chance of occurrence of mismatch answers. The ESS questions can be categorized by means of more characteristics than show cards and our own distinction of conversational character, difficulty and ambiguity. In the next section we will describe the relation between other question characteristics and the percentage of mismatch answers in Q-A sequences.

6.5 Different structural types of questions and the occurrence of mismatch answers

The 268 ESS survey questions were divided into several question types by means of structural characteristics, which were the question proper (assertion, choice question, yes-no question or open question) and the way the alternatives for these questions are presented. Alternatives can be implied by the question with an open response format (i.e., numbers or percentages, or 'yes' and 'no' in case of yes-no questions). Alternatives can also be used in a field-coded question format (a list of alternatives that is not literally implied by the question format but also not read by the interviewer with initial question reading). Furthermore alternatives can be read within the question or afterwards, or they can be presented on a show card.

In Table 6-18 the different question types, and the number of questions concerned in the ESS data are shown. The table shows the types of alternatives used (as shown in the columns) and the type of question proper (as shown in the rows). For a large number of questions, show cards were used, and show cards were not equally distributed over the question types. Show cards were systematically used for specific types of questions, namely assertions.

Table 6-18 Different types of questions in the ESS

	Alternatives:				
	Implicit	Field coded	Read within	Read after	On show card
Assertion	-	-	-	-	39
Closed Q	22	1	7	17	107
Yes-No Q	44	3	-	3	-
Open Q	7	-	-	1	14

6.5.1 Occurrence of mismatch answers and the use of show cards

Table 6-19 shows that there is no difference in the percentage of Q-A sequences with a mismatch answer between the questions with and without show cards.

Table 6-19 Percentage of Q-A sequences with mismatch answers for questions with and without show cards

	Questions without show cards		Questions with show cards		Total	
Q-A sequence	Freq.	%	Freq.	%	Freq.	%
With mismatch answer	148	12%	309	13%	457	13%
Without mismatch answer	1113	88%	2053	87%	3166	87%
Total	1261	100%	2362	100%	3623	100%

χ^2 = n.s.

Other studies (Dijkstra and Ongena forthcoming; Prüfer and Rexroth 1985), have shown that questions with show cards yield significantly *less* problematic Q-A sequences than questions without show cards. This difference is especially due to a high number of mismatch answers for questions without show cards.

In the ESS data this difference seems to be absent. Apparently, show cards were prescribed for the right questions. They were left out for questions that yield the lowest number of mismatch answers anyway (for example yes-no questions), and were used for actually the most problematic questions, so that, despite of the benefit of show cards, these questions still yield a large amount of mismatch answers. It is of course also possible that show cards for different questions were mixed up during the interview and as a result respondents gave mismatch answers.

Although this could not be confirmed for the ESS data, it is likely that show cards indeed make the respondent's task easier. It certainly makes sense that the visual presentation of response alternatives lowers the chance of the occurrence of mismatch answers. Only a split ballot experiment in which the same questions with and without show cards are compared can demonstrate this effect of show cards. In such an experiment, the conversational character of alternatives could also be manipulated for show cards. In this way the effects of both alternatives and show cards can be tested systematically.

The use of show cards affected the course of Q-A sequences for questions *without* show cards: it was not always clear to respondents that no show card was available for a particular question. Respondents who still had a show card in front of them sometimes tried to answer a non-show card question by means of this show card, as is illustrated in Excerpt 6-4. In this case, the show card of the preceding question presented a ten-point scale for 0 (very dissatisfied) to 10 (very satisfied). Such confusion caused some interviewers, in their later interviews, to explicitly mention the alternatives for yes-no questions before the respondent's answer. Confusion concerning show cards may have been caused by the frequent change in response format: 55 different show cards were used for the 160 show card questions (i.e., show cards were changed every other three to four questions).

Excerpt 6-4 Q-A sequence concerning ESS question K18

1. I: During the last 12 months, have you made any attempt to improve conditions at work, or to prevent them from getting worse?
2. R: Seven
3. I: Yes or no?
4. I: Sorry
5. R: Hihi hihi
6. R: I already thought what a strange question
7. R: Yes
8. I: Yes?
9. R: Yes

6.5.2 Occurrence of mismatch answers for show card questions

Table 6-20 shows the percentage of mismatch answers occurring in the Q-A sequences concerning the three different types of questions with show cards. In case of improper open questions, the question was posed as an open question, but respondents had to choose from a list of options. These questions appeared to yield a large percentage of Q-A sequences with a mismatch answer.

Table 6-20 Percentage of mismatch answers for three different types of questions with show cards

Q-A sequences	Assertions		Closed choice question		(Improper) open question	
With a mismatch answer	94	14%	170	11%	45	29%
Without a mismatch answer	597	86%	1345	89%	111	71%
	691	100%	1515	100%	156	100%

$\chi^2 = 38.87$ df = 2, $p < 0.01$

Most improper open questions in the ESS concerned questions that asked for participation in organizations. Excerpt 6-5 shows an interaction with such a question. It was possible to select multiple options. However, respondents frequently just said ‘yes’ or ‘no’, as is also the case in Excerpt 6-5. Furthermore, respondents had difficulties with the distinction of response alternatives; for example, being member of an organization usually implies donating money.

Excerpt 6-5 Q-A sequence concerning ESS question K10.11b

1. I: What have you done for the following organization umh a society for cultural or hobby related activities?
2. R: Yes
3. I: Have you been a member or eh
4. R: I am at the moment uh taking a computer course

Listed alternatives for ESS question K10.11b

1 Member
2 Participated
3 Donated money
4 Voluntary work

0 None of these four

6.5.3 Occurrence of mismatch answers and presentation of alternatives without show cards

Sections A and B of Table 6-21 show the percentage of mismatch answers occurring in the Q-A sequences concerning choice questions and yes-no questions, for the different types of presentation of response alternatives. When we compare sections A and B, it can be seen that choice questions with implicit alternatives yield more mismatch answers (12%) than yes-no questions with implicit alternatives (i.e., 1%, χ^2= 82.92, df = 1, p < 0.01). This finding makes sense, as choice questions with implicit alternative were mostly difficult questions (i.e., 70%), whereas yes-no questions were mostly easy questions (86%). For the other types of alternatives, there is no difference between the questions in the percentage of mismatch answers.

Table 6-21 Percentage of mismatch answers for different types of presentation of alternatives

A. Choice Questions	Implicit		Field coded		Read within Q		Read after Q	
Q-A sequences with mismatch	51	12%	1	13%	14	9%	48	20%
Q-A sequences without mismatch	226	82%	7	87%	84	91%	191	80%
Total	277	100%	8	100%	98	100%	239	100%
	χ^2= n.s.				χ^2= n.s.			
B. Yes-No questions	Implicit		Field coded		Field coded		Read after Q	
Q-A sequences with mismatch	6	1%	9	17%	9	17%	14	39%
Q-A sequences without mismatch	525	99%	42	83%	42	83%	22	61%
Total	531	100%	51	100%	51	100%	36	100%
	χ^2= 50.57, df = 1, p < 0.01				χ^2= 4.89, df =1, p < 0.05			

Within yes-no questions, there are significant differences in the percentage of mismatch answers between the types of presentation of alternatives. Especially when a question is worded as a yes-no question, but nevertheless forces the respondent to choose from a list of alternatives that consist of alternatives other than 'yes' and 'no', this question yields a high number of mismatch answers.

Mismatch answers occur most frequently in Q-A sequences concerning yes-no questions with listed alternatives that are read *after* the question. This is not surprising; as the alternatives are read after the question, the respondent is likely to interrupt the interviewer, before all response alternatives are read. This interruption is avoided when alternatives are read within the question, i.e., before the question delivery component had been read (Houtkoop-Steenstra 2000).

Also field-coded yes-no questions (with listed instead of implicit alternatives, which interviewers are not required to read) obtain a high percentage of mismatch answers. This clearly indicates the problem of field coded questions, as is illustrated by Excerpt 6-6. Although the question is formulated as a yes-no question, the response alternatives that are listed included not only 'yes' and 'no', but also specifications of 'yes'. These specifications were not indicated by the question, and as a result the question relies on good interviewer behavior such as a non-directive probe for this specification in case the respondent answers the question with 'yes'.

Excerpt 6-6 Q-A sequence concerning ESS question F8

1. I: Are you hampered in your daily activities in any way by any longstanding illness or disability, infirmity or mental health problem?
2. R:Yes
3. I: So, is that a lot or to some extent?

Listed alternatives for ESS question F8:
1 Yes, a lot
2 Yes, to some extent
3 No
8 (Don't know)

6.6 Conclusion

For questions without show cards the results indicated confirmation of hypotheses 1 and 2a. Conversational questions yielded more mismatch answers than formal questions, and formal alternatives yielded more mismatch answers than conversational alternatives. However, some of these variables are confounded with other characteristics of questions. This makes interpretation of results rather difficult. For example, assertions occurred only as questions with show cards and formal alternatives in the ESS questionnaire. As we also showed in chapter 5, assertions typically yield a high number of mismatch answers, but show cards reduce the chance of mismatch answers. Apparently, the survey designers of the ESS have

chosen the right questions to be accompanied by show cards. However, from our analyses we cannot conclude that show cards were effective, because the questions with show cards still yielded an equally high percentage of mismatch answers as compared to questions without show cards.

The results did not confirm hypothesis 2b (listed alternatives yield more mismatch answers than implicit alternatives). Questions with implicit alternatives generated a high percentage of mismatch answers, probably due to the fact that they mostly concerned difficult questions. We could confirm hypothesis 2b, only considering the difference between implicit and formal alternatives, for easy questions. Questions that account for this effect were questions such as the year of birth of the respondent, which yielded no mismatch answers at all. This question not only correctly implies the response alternatives (i.e., a year), it also asks for information that is easily found in memory.

The judgment of categorizing the three types of mismatch answers can be considered rather subjective. Especially the fact that mismatch answers that were not recognized as task or as cognitive mismatch answers were all categorized as conversational mismatch answers is dubious. However, it was useful to relate the three types of mismatch answers to the different types of questions. Conversational mismatch answers occurred most frequently for conversational questions with formal alternatives. This indicates that such questions can evoke a conversational style. Respondents may have the goal to get most benefits out of the interview, at as low possible costs. They may want to have a pleasant conversation, with little cognitive effort, and at the same time present themselves as cooperative persons.

Furthermore, the results showed that it is important to keep the response format as constant as possible. In the ESS data, the response format was most frequently changed with the use of many different show cards. Such changes in response format confused respondents. This confusion urged interviewers to notify the respondent of the required answering format, even in case of yes-no questions. It is even more probable that a frequent change in response format will confuse respondents in surveys without show cards.

Task mismatch answers typically occur when respondents are uncertain about the meaning of definitions of concepts in questions, or when they have problems with choosing the right response alternative. Task mismatch answers indicate that respondents approach the interview in a more serious, task-oriented way. This also appears from the (indirect) requests for clarification that often precede task mismatch answers.

Cognitive mismatch answers arise because of problems with the availability of the information required by the question. These mismatch answers may indicate retrieval and judgment strategies that respondents use to arrive at an answer, which may consist of enumerations or estimations (e.g., retrieved rates of occurrence instead of individual instances).

Although a large number of different questions, and as a result a relatively large number of Q-A sequences was available, data from only twenty-three different respondents and seven interviewers were analyzed. Because of these low numbers, the results cannot

easily be generalized to other surveys, and it does not make much sense to analyze differences between respondents (as was done for the data analyzed in chapter 5).

To test the hypotheses, we used ad hoc categorizations of question wording and types of alternatives. The categories appeared to be correlated, which complicated the interpretation of the results. Especially effects for questions with show cards versus questions without show cards are difficult to interpret. Furthermore, non-ambiguous questions were rated as formal questions more frequently than ambiguous questions. Thus, comparing ambiguous and non-ambiguous questions was highly correlated with a comparison of the conversational character of the questions. Our results indicated that for difficult questions, conversational question wording increases the chance of mismatch answers as compared to formal question wording, but for easy questions this difference was not found. Furthermore, effects of the conversational character of questions seemed to be different for perceptual, factual and opinion questions. In order to further test such effects we need to compare different versions of the same question.

Finally, distinctions of the conversational character of questions and types of alternatives were based upon the researcher's assumptions about common words used in ordinary conversations. A better strategy would be to use empirical data concerning words and formulations used in ordinary conversations of the intended population (i.e., the Dutch population), in order to formulate new conversational and formal questions and alternatives. In an experimental design, the same questions can be manipulated, and confounding of variables can be avoided. In the next chapter a study with such operationalizations, to test the same hypotheses in an experimental design, is described.

7 An experimental study on question wording

7.1 Introduction

In the non-experimental study with the European Social Survey data (ESS) described in the previous chapter, we tested effects of different question types that were not manipulated systematically. However, the wording of questions was in some cases confounded with other relevant question characteristics. The clearest example of such confounded variables are the topic of the questions and their wording. For example, the background questions (respondent's age, number of persons in household) were all categorized as formal questions. Another example of a confounded variable is the specific question type; all assertions were accompanied by formal alternatives. Furthermore, conversational and implicit alternatives were only found for formal questions, but not for conversational questions. Such a confounding prevents finding effects of the intended question characteristics (question wording and types of alternatives) independent from other question characteristics (topic or type of question).

From these issues we may conclude that we need an experimental study, in which the same questions are systematically varied for question wording and the types of alternatives to test the hypotheses, in order to warrant a better internal validity of the study.

Another problem of the ESS-data was that they concerned face-to-face interviews, whereas our hypotheses resulted from the analysis of telephone interviews (i.e., the television survey described in chapter 5). In face-to-face interviews, non-verbal communication can play an important role in the interaction. For example, specific task-related behaviors, such as acknowledgements or confirming responses like a nod or specific gesture may occur, without auditory communication. This non-verbal visual communication between interviewer and respondent is not available to the researcher, when only auditory information is used. In case of telephone interviews, there is no visual communication and as a result audio-recordings are sufficient to perform interaction analysis.

Finally, telephone interviewing enables conducting a large number of interviews in a relatively short period of time. For these reasons, it was decided to create a full experimental design and to conduct the interviews by telephone (CATI).

In the next sections the experimental design, the topic of the questionnaire and the operationalizations of the manipulations will be described. Next, the procedures in conducting the interviews, the response rates, and coding procedures will be presented. Finally, the results of the analyses, testing the hypotheses, will be described, followed by a discussion.

7.1.1 Experimental Design

The experiment concerns five different hypotheses, (i.e., the same as described in section 6.2):

H1 *A question that is formulated as a conversational question will generate more mismatch answers, than a formally worded question.*

H2a *Questions with conversational response alternatives will generate less mismatch answers, than the same questions with comparable formal response alternatives.*

H2b *Questions with listed (i.e., conversational or formal) response alternatives will generate more mismatch answers than questions with implicit response alternatives*

H3 *Questions requiring information not readily available in memory (i.e., difficult questions) will generate more mismatch answers, than questions requiring relatively little cognitive processing (i.e., easy questions).*

H4 *Questions containing ambiguous concepts will generate more mismatch answers than questions not containing ambiguous concepts.*

To test the hypotheses concerning the effects of question wording on the occurrence of mismatch answers, it was necessary to create multiple conditions with different question wordings. The experimental manipulations were varied within subjects; respondents were asked a number of conversational as well as formal questions, easy questions as well as difficult questions, and questions with ambiguous concepts as well as questions with less ambiguous concepts. However, we preferred to manipulate wordings of the same questions without confronting respondents with different wordings of the same questions in one interview. Consequently, different types of the same question were compared between respondents, in a split-ballot design.

Table 7-1 shows the different combinations of questions. As the table shows, not all possible combinations were included. We could not use to many different questions, because a telephone survey is limited in its length. Thus we selected the most important and realistic combinations in order to avoid the effects of many low cell values in the data. The excluded combinations are indicated in the table with a hyphen (-). The included combinations were focused on the main hypotheses, which concern comparisons of formal and conversational wording of questions and alternatives.

Table 7-1 Question-combinations used in the experiment

Wording of Questions	*Wording of response alternatives*		
	Formal alternatives	Conversational alternatives	Implicit alternatives
-Formal -Easy -Non-ambiguous Q	FFEN[1]	CFEN	IFEN
-Formal -Easy -Ambiguous Q	-	CFEA	IFEA
-Formal -Difficult -Non-ambiguous Q	FFDN	-	IFDN
-Formal -Difficult -Ambiguous Q	-	-	IFDA
-Conversational -Easy -Non-ambiguous Q	FCEN	CCEN	IFEN
-Conversational -Easy -Ambiguous Q	-	CCEA	ICEA
-Conversational -Difficult -Non-ambiguous Q	-	-	ICDN
-Conversational -Difficult -Ambiguous Q	-	-	ICDA

The first digit in the abbreviation refers to the columns; the next three digits refer to row values. The abbreviations used indicate the combination of the four manipulations. These are: the conversational character of alternatives (F, C or I; i.e., Formal, Conversational or Implicit), conversational character of questions (F or C; i.e., Formal or Conversational), difficulty (E or D; i.e., easy or difficult) and ambiguity (A or N; i.e., ambiguous or non-ambiguous) respectively. Hence, 'FFEN' means Formal alternatives, Formal, Easy, and Non-ambiguous question.

Testing the hypotheses will involve several comparisons of different conditions. For example, to test hypothesis 1, we compare formal questions with conversational questions, and keep the other manipulations constant. For example, type FFEN (see explanation below Table 7-1) and FCEN are both non-ambiguous and easy questions with formal alternatives, but they differ with respect to the conversation-likeness of their question wording (i.e., formal and conversational respectively).

7.1.2 Composing the questionnaire

An important requirement was the use of a questionnaire that allowed us to maximize the external validity of the study. Ideally, we aimed to obtain results that can be generalized to actual surveys. Therefore we tried to use the original question wordings of questions from actual surveys, and established for each question the condition to which the particular question belonged (i.e., whether the question could be considered as conversational or formal, easy or difficult, ambiguous or non-ambiguous). Subsequently, we reworded the question

according to the 'opposite' conditions. In addition, we took care of variation of question types, e.g., we included assertions that concerned perceptions or opinions as well as choice questions that concerned behaviors, facts or perceptions.

An alternative strategy is to write new questions for all conditions, without necessarily using the original question wording as derived from an actual survey in any of the conditions. Such a strategy contributes positively to the internal validity of the experiment, but negatively to external validity. For several questions we used a strategy that takes advantage of both strategies mentioned above: we took question wordings of surveys and used both conversational and formal aspects of those questions for new wordings in the two conditions.

7.1.3 Topic of the questionnaire

The topic of the questionnaire that was designed for the experiment was health and health-related issues. As König-Zahn, Furer and Tax (1993) describe, in the field of health measurement an enormous amount of questionnaires exist. In their comparison of health questionnaires they restricted themselves to general health concepts. Furthermore, they excluded questionnaires that were lacking methodological accounts; that were used in only one study, or that were designed for particular age groups or specific populations. These restrictions completely fulfill our requirements, and we decided to use some of the questions from the questionnaires they evaluated (for example the General Health Perception Questionnaire and the Statistics Netherlands Health questionnaire).

As health surveys are very frequently conducted, it is no surprise that many behavior coding studies were also applied to health surveys (Cannell et al. 1968; Dykema et al. 1997; Hill and Lepkowski 1996; Lepkowski et al. 2000; Mathiowetz and Cannell 1980; Oksenberg et al. 1991; Presser and Blair 1994; Snijkers 2002). Most of the survey-questions that were analyzed in these behavior coding studies concerned questions about medical consumption (doctor and hospital visits etc.).

In Appendix 7-1, the question wording of all questions used in the manipulated questionnaires is given. In section 7.2, the operationalizations of the manipulations are presented.

7.2 Operationalizations

7.2.1 Operationalization 'conversational' versus 'formal' questions

Hypothesis 1 concerns a difference in 'conversational' and 'formal' questions. As discussed in chapter 2, there are several aspects of 'ordinary' conversations that can be contrasted with standardized survey interviews. For example, in ordinary conversations speakers have many possibilities to clarify misunderstandings. In standardized interviews, however, interviewers are often prohibited to clarify question wording. This creates standardization of input instead of standardization of meaning (Schober and Conrad, 2002). Furthermore, in ordinary conversations utterances are adapted to specific recipients and situations. In standardized

interviews, interviewers are not allowed to change question wording, for instance by adapting it to respondents that spontaneously gave the information earlier.

Focusing on a different aspect of ordinary conversations, our definition of 'conversational questions' primarily concerns the word choice in ordinary conversations. Words that are used in surveys may differ from those typically used in ordinary conversations. For example, in survey questions, 'research-theoretical concepts' may be used that would be awkward to use in ordinary conversations, such as 'main activity' in a common survey question like 'What is currently your main activity: employed, unemployed, retired, or in education?' Adhering to 'lexical availability' and 'frequency of use' factors (Brennan and Clark 1996), in 'conversational' questions we will use words that are used most frequently in ordinary conversation, but will have more or less the same meaning as their formal equivalents. The conversational questions are compared with 'formal' questions. Formal questions contain words that are less frequently used in ordinary conversations.

Of course it is not possible to choose conversational words for all elements of a conversational question, and only formal words for a formal question. The difference in word use especially concerns the *nouns*, *verbs, adjectives* and *adverbs* used. Although other word types (i.e., articles, conjunctions, pronouns, prepositions) and the grammatical structure used may differ between conversations and survey interviews, we will not address this in our manipulations.

To determine what words are used most frequently we used information from the Spoken Dutch Corpus project. This project aims to yield a resource of 1,000 hours of speech (approximately ten million words) originating from adult speakers of standard Dutch (Oostdijk et al. 2002). We used release 6 (November 2002), which contained for 143,851 different words a total frequency of 6,023,935 of word occasions. From the available frequency tables we deleted entries for punctuation marks (i.e., comma's, dots, question marks were not counted as words).

Within the corpus, two different sub-corpuses are relevant as indicators for ordinary conversations, i.e., the 'spontaneous conversations' and 'telephone conversations'. These sub-corpuses contain 1,733,244 and 593,980 words respectively. The total corpus also includes sub-corpuses that were not considered as relevant for ordinary conversations, such as read aloud written texts and radio broadcasts.

The strategy to select words appropriate for question wording by frequency of word occurrence was first described by Payne (1951). He used *The Teacher's Word Book of 30,000 Words,* which was based upon a count of 4,500,000 words in popular magazines that appeared between 1927 and 1938. However, Payne advises to use the word frequency in a negative way, i.e., abandoning words of low frequency from a list of synonyms, instead of choosing the words with the highest frequency. He argues that "word count taken alone is not a sufficient guide for the selection of words. It is useful for the elimination of low frequency words, but affords no guarantee of perfection in high-frequency words" (Payne 1951, p. 145). Payne warns that different meanings of words (in different contexts or with different pronunciations) are not incorporated in the word frequencies. Furthermore, compound words,

as 'public opinion', are counted as isolated words, i.e., 'public' and 'opinion', whose separate descriptions cannot clearly convey the meaning of the combinatory concept.

As compared to Payne's source of word frequencies, the frequency database of the Spoken Dutch Corpus project probably is more appropriate. Firstly, it consists of words in actual ordinary conversations. Secondly, it is of course also more recent. Thirdly, we were obliged to use a corpus of Dutch words, as the interviews were held in The Netherlands.

Nevertheless, the word frequency database does not give separate frequencies for all possible meanings of the same word. The same word that is frequently used with a specific meaning could only be seldom used with a different meaning. For example, the Dutch word 'eens' can have the meaning 'once (upon a time)' or 'some day', but it can also have the meaning 'agreed'. The corpus however only gives frequencies for the word 'eens' without specifying for which meaning(s). This creates a problem if we use a word with a high frequency and define it as 'conversational', whereas this high frequency is caused by its meaning different from the one we use it for the formulations of our questions. This problem can be dealt with by means of checking the number of word meanings and their arrangement listed in a dictionary. Arrangement of several meanings of a word is often in accordance with their frequency of use in a language, with the most common meaning given first, although historical ordering is also used. We use the criterion that high frequency words with multiple meanings can only be conversational when they are in the first half of all meanings listed for a word in the dictionary. Once we have determined the difference in frequency for two synonyms with a specific meaning in ordinary conversations, in order to determine which one should be considered as conversational and which one as formal, we have another problem. According to the Principle of Contrast (Clark, 1987, cited in Brennan, forthcoming) "there is no such thing as a true synonym; even if two words seem interchangeable in a particular context, there are other contexts in which they contrast" (Brennan forthcoming, p. 7). However, Schuman and Presser (1981) conclude, based on results of their experiments on question-wording, that they cannot accept the notion that any change in its wording will result in an entirely new question. We therefore assume that the use of synonyms will not have problematic consequences for the meaning of questions, and consequently for our interpretation of the effects of question wording on the occurrence of mismatch answers.

The following definitions will be used:

Conversational questions are questions in which nouns, verbs, adjectives and adverbs consist of words that are common in conversations (as indicated by high frequencies in the Spoken Dutch Corpus). In the Corpus, the average frequency of words is the total frequency of words divided by the number of different words, which was 34 in the 'spontaneous conversations', and 26 in the 'telephone conversations'. Thus, we define a 'high' frequency as a frequency that is above the mean frequency of 34 the sub-corpus 'spontaneous conversations' and above the mean frequency of 26 of the sub-corpus 'telephone conversations'. Furthermore, when a

certain word has different meanings, the meaning as used in the survey question should be in the *first* half of all meanings listed in the Dutch Dictionary.

Formal questions are questions in which nouns, verbs, adjectives and adverbs consist of words that are not common in conversations (as indicated by low frequencies in the Spoken Dutch Corpus). A 'low' frequency is a frequency that is below the average frequency of 34 of the sub-corpus 'spontaneous conversations' and below the average frequency of 26 of the sub-corpus 'telephone conversations'.

As indicated in section 7.1.2, we selected, as far as possible, existing questions from actual surveys, and determined by means of the above mentioned criteria whether the questions could be considered as conversational or formal. In a few cases the original question was slightly adapted to fit more accurately in the category that it was assigned to. Subsequently, with replacement of words (i.e., conversational words with formal words and vice versa), we constructed an alternative question (i.e., the opposite version). We took care that the frequency of the relevant words in the conversational version was at least five times as large as in its formal counterpart.

7.2.2 Some examples of conversational and formal questions

In the questionnaire, 40 questions were manipulated for the conversational character of the question. The exact wording of all questions that were used can be found in Appendix 7-1. Specific questions will be referred to by their question number in this appendix.

We derived question wordings for 26 of the 40 questions from actual surveys. Of those 26 questions, for 12 questions we used literal question wording for one of the conditions, and formulated our own 'opposite' condition, and for 11 questions we used adapted question wording for both conditions. For the remaining three questions (which were background questions) we found both a conversational version and a formal version in existing surveys, thus we could use literal wording for both conditions. Below we will give some examples of question wordings that were used in the questionnaire. The examples are presented in separate sections for three batteries of assertions and choice questions.

General Health Perception assertions

In the questionnaire eight questions (Q2-Q9) were derived from the General Health Perception Questionnaire (Brook et al. 1979). In the GHPQ the subjective perception of one's own health status, in the past, at present and as expected in the future is measured. The questions are formulated as assertions. Two Dutch translations of the short version (SF-20) of this questionnaire were available (Kempen et al. 1995; Kriegsman, Van Eijk and Deeg 1995). Three examples of assertions in original wording, and the assertions reworded as formal questions, are listed in Table 7-2. As might be observed by face value in this table, the original GHPQ-wording was typically conversational. However, for the English translations of the questions, which are included for illustration purposes, we did not use word frequency

data. As a consequence, the difference may be less apparent as in their Dutch equivalents.

Table 7-2 Original conversational wording and formal rewording of GHPQ SF-20 assertions (Dutch wording in italics)

In original (conversational) wording	Reworded as a formal question
I worry about my health[1] *Ik maak me zorgen over mijn gezondheid*	My health causes me worries *Mijn gezondheid baart mij zorgen*
I believe that sometimes I am just going to be sick *Ik accepteer het dat ik soms gewoon ziek wordt*	I acknowledge that from time to time I will become sick *Ik aanvaard dat ik van tijd tot tijd nu eenmaal ziek wordt*
I was so sick once I thought I might die *Ik ben wel eens zo ziek geweest dat ik dacht dat ik doodging*	I have on one occasion been so sick I thought I would pass away *Ik ben een enkele maal zo ziek geweest dat ik dacht te overlijden*

[1]From the original wording "I never worry about my health" the word 'never' was omitted because it is obviously problematic to use negations in question wording.

Public health assertions

In the questionnaire, six assertions concerning the respondent's opinion on costs of public health were included (Q27-Q32). The questions were derived from Dutch and Belgian studies (Bernts 1991; Elchardus, Tresignie and Derks undated; Van de Berg, Jansen and Haveman 1986). To illustrate how we reworded the original formal questions into conversational questions, both wordings are presented in Table 7-3, for the first assertion within this battery. Most questions were originally worded as formal questions. The first four questions had a similar question format, but differed in one word, indicating a different concept. For example, in the second assertion as compared to the first assertion, the word 'smoking' is replaced by 'alcohol'. A difference with the GHPQ assertions is that the public health assertions concern the respondent's opinion on a social issue, whereas the GHPQ assertions concerned the respondent's evaluation of one's own health, i.e., concerned a perceptual judgment question.

Table 7-3 Original conversational wording and formal rewording of public health assertions (Dutch wording in italics)

Original (formal) wording	Reworded as a conversational question
Somebody is responsible for the extra costs of public health caused by smoking *Iemand is zelf aansprakelijk voor de extra ziektekosten door roken*	You should pay for your owns costs of public health caused by smoking *De extra ziektekosten door roken moet je zelf betalen*

Government and health assertions

In the questionnaire, three assertions (Q18-Q20) taken from the Dutch Government 'Belevingsmonitor' ('Perception monitor' RVD 2003) were included. These concerned the respondent's opinion about a smoke free hotel and catering industry, education about

smoking, and inspection of food safety. For these assertions we used the original formal wording and a conversational rewording.

Choice questions

In the questionnaire, assertions were alternated with choice questions concerning factual, behavioral and perceptual information. We included factual questions concerning the respondent's age, employment status, level of education etc. Furthermore, we included behavioral questions, such as the number of days that respondents watch television and use breakfast. We also included questions asking respondents to rate their health and weight. Such questions can be labeled as 'perceptual' questions (Kalton and Schuman, 1982), as they concern respondents' estimations of aspects of their own health, and do not concern facts, behaviors or opinions. Of the 25 choice questions, 24 were manipulated for the conversational character of the question. For four choice questions we used a question wording that was literally derived from an actual survey. For three of these questions (Q35, body length, Q39, employment and Q41, age) we could use literal question wording for both the formal and the conversational version, as both versions appeared in different surveys. For six questions we transformed the original questions that contained both conversational and formal words into more extreme versions of conversational or formal questions. For the remaining fourteen questions we created entirely new questions for both conditions.

The manipulation that we used for questions with implicit alternatives comprised a difference in the extent to which the alternatives are explicitly implied by the question wording. For the conversational questions we used the words "how many" to imply the alternatives (i.e., a number of days/hours etc.). For the formal questions we used the words "What is the number of" to imply the same alternatives. The latter can be considered as a wording that is not commonly used in ordinary conversations, but more explicitly implies the required response alternatives. In Table 7-4 we present examples of a factual, behavioral and perceptual choice question respectively.

Table 7-4 Formal and conversational wording of choice questions (Dutch wording in italics)

Formal wording	*Conversational wording*
What is your year of birth? *Wat is uw geboortejaar?*	How old are you? *Hoe oud bent u?*
What is the number of days a week that you usually watch television? *Wat is het aantal dagen per week, dat u doorgaans televisie kijkt?*	How many days a week do you usually watch TV? *Hoeveel dagen per week kijkt u meestal TV?*
Do you find yourself too light, too heavy or do you consider your weight as not too light and not too heavy? *Vindt u zichzelf te licht, te zwaar of vindt u uw gewicht niet te licht en niet te zwaar?*	Do you consider yourself too skinny, too fat or do you think your weight is good? *Beschouwt u zichzelf te mager of te dik of beschouwt u uw gewicht als goed?*

7.2.3 Operationalization 'conversational' versus 'formal' alternatives

Hypothesis 2a concerned the difference between conversational and formal alternatives. Corresponding with our definitions of conversational questions, our definition of conversational alternatives concerns the *word choice* in ordinary conversations.

The fact that for answers produced in conversations some words are preferred over others can be illustrated by the difficulty people experience when they are forced to avoid conversational words, which was the essence of a game on a popular Dutch children's radio show in the eighties ('Ko de Boswachtershow'). In this game participants had to talk with the host of the radio show, but they were not allowed to say the words 'yes' and 'no', whereas they were asked yes-no questions continuously.

It is important to avoid that questions evoke a large number of mismatch answers for other reasons than the manipulation of answer alternatives. The structure of a question may evoke premature answers (i.e. the respondent answers before the interviewer has read the answer alternatives, see Stax 2004). These premature answers are likely to emerge as mismatch answers, because the respondent has not been informed about the response alternatives that are used. Such an effect may rule out effects of the manipulated response alternatives, because the structure of the question caused mismatch answers irrespective of the type of alternatives. Houtkoop-Steenstra (2000) suggests to incorporate the answer alternatives within (or to let them precede) the question. In this way the question component is delivered last. This assures that respondents are not able to infer the meaning of the question before all alternatives are read, and the chance of premature answers is reduced.

In Example 7-1 we give a question wording that is likely to generate a premature answer, i.e., question (a), and a question wording that is less likely to do so, i.e., question (b). Hence, both questions with conversational answer alternatives and questions with formal answer alternatives were formulated consistent with the format of question (b): the question component is last delivered and the answer alternatives are incorporated within the question. However, this structure may create grammatical problems in some cases, especially in the English language. Therefore it is not always possible to present our Dutch question wording in English equivalents. Furthermore, for the assertions a different structure was necessary. For all assertions within a battery the same alternatives were used. The alternatives were read only at the introduction of a new battery of assertions. As our analyses of the Television survey showed (see section 5.5, chapter 5), not repeating the alternatives for individual assertions in a battery increased the chance of mismatch answers. However, this concerned the use of four or five different alternatives. The number of alternatives used is expected to affect the chance of mismatch answers. If more than five alternatives are orally presented, respondents are not likely to remember all alternatives. Therefore, a large number of alternatives forces survey researchers to use show cards. As we indicated in chapter 6 (section 6.3.4) effects of alternatives for show card questions are difficult to interpret. Therefore, in the present experiment, for most assertions only two or three alternatives were used. We expect that respondents would not have too much difficulty in remembering two or

three alternatives. Interviewers were of course instructed to repeat the alternatives if respondents did not provide an adequate answer.

Example 7-1, Places of the question component

(a)	How much have you heard about X? A great deal, some, not very much or nothing at all?
(b)	'Have you heard a great deal, some, not very much, or nothing at all about X?

7.2.4 Some examples of conversational and formal alternatives

For the GHPQ assertions, three response alternatives were used. These alternatives were worded either conversationally ('yes', 'maybe' or 'no') or formally ('true', 'possibly true' and 'false'). According to both relevant sub-corpuses of the Spoken Dutch Corpus, the average frequency of the conversational alternatives was twenty times as high as the average frequency of the formal alternatives.

The Public Health assertions were accompanied by two answer alternatives, either conversational ('yes' or 'no') or formal ('agree' or 'disagree'). The average frequency of 'yes' and 'no' was 96 times higher than 'agree' and 'disagree'.

For the 'Government and Health' assertions we used conversational alternatives 'yes' and 'no' versus formal alternatives with a 5-point scale ranging from 'strongly agree' to 'strongly disagree'. This in fact is an uneven comparison, as not only the wording of the alternatives differs, but also the number of alternatives used. However, in this case we aimed to compare the most extreme versions of the two types of alternatives: the conversational alternatives as simple (and as optimal) as possible, and the formal alternatives as used most frequently in common surveys. The ratio in average Corpus word frequency for conversational alternatives versus formal alternatives was 144:1. In Table 7-5 the conversational and formal alternatives for the three batteries of assertions are presented. To illustrate the context of the alternatives, the introductory statements that belong to the answer alternatives are included.

Table 7-5 Wording of conversational and formal answer alternatives for the assertions

Conversational answer alternatives	*Formal answer alternatives*
GHPQ-Assertions:	
Now I am going to read some assertions. To indicate whether you agree or disagree with the assertion you can answer with *'yes', 'maybe' or 'no'*.	Now I am going to read some assertions. To indicate whether you agree or disagree with the assertion you can answer with *'true', 'possibly true' or 'false'*.
Public health assertions:	
Now I am going to read some assertions. To indicate whether you agree or disagree with the assertion you can answer with *'yes' or 'no'*.	Now I am going to read some assertions. To indicate whether you agree or disagree with the assertion you can answer with *'I agree' or 'I disagree'*.
Government and health assertions:	
Now I am going to read some assertions. To indicate whether you agree or disagree with the assertion you can answer with *'yes' or 'no'*.	Now I am going to read some assertions. To indicate whether you agree or disagree with the assertion you can answer with the following five answer possibilities *'strongly agree, agree, neutral, disagree or strongly disagree'*.

For a subset of the *choice questions* we also manipulated the wording of the answer alternatives. Some examples of formal and conversational wordings of response alternatives for behavioral questions are shown in Table 7-6. For the wording of the alternatives we did not only use conversational or formal alternatives, but also implicit alternatives. Response alternatives are implicit when they are not explicitly listed and the question gives an indication of what kind of answer alternatives are required. Questions with implicit alternatives typically begin with phrases like 'how many' and 'how often'. Furthermore, a number of factual questions were manipulated in addition in order to test hypotheses 3 and 4, concerning the difficulty and ambiguity of the questions (see section 7.2.6).

Table 7-6 Original conversational wording and formal rewording of choice questions (Dutch wording in italics)

Formal wording	*Conversational wording*
Do you usually walk at *slow, normal or fast* pace? *Wandelt of loopt u in langzaam, gewoon of snel tempo?*	Do you usually walk at *low, average or high* pace? *Wandelt of loopt u op een laag, middelmatig of hoog tempo?*
Do you practically watch 0 days, 1 to 3 days, 4 to 6 days or 7 days a week TV? *Kijkt u praktisch 0 dagen, 1 tot 3 dagen, 4 tot 6 dagen of 7 dagen per week TV?*	Do you watch television every day, most days, some days or hardly ever? *Kijkt u elke dag, de meeste dagen, sommige dagen, of bijna nooit TV?*
Do you on weekdays never, 1 to 2 days, 3 to 4 days or all 5 days use corn products such as bread, muesli or cornflakes as breakfast*?* *Gebruikt u doordeweeks nooit, 1 tot 2 dagen per week, 3 tot 4 dagen per week of alle 5 dagen graanproducten zoals brood, muesli of cornflakes bij het ontbijt?*	Do you on weekdays never, 1 to 2 days, 3 to 4 days or all 5 days use corn products such as bread, muesli or cornflakes as breakfast? *Gebruikt u doordeweeks nooit, af en toe, de meeste dagen, of elke dag graanproducten zoals brood, muesli of cornflakes bij het ontbijt?*

7.2.5 *Operationalization easy versus difficult questions*

Hypothesis 3 concerns the difference between questions requiring little cognitive effort (easy questions) and questions requiring a relatively high amount of cognitive effort to answer (difficult questions). A question will require more cognitive effort when it concerns distant events instead of recent events, or when it requires computing in large steps rather than small steps. Computation in small steps is done by means of decomposition. Generally, questions are decomposed by means of taking shorter reference periods (e.g., a week instead of a month) or by means of asking separate questions for specific categories (e.g., separate questions for consumption of beer, wine, liquor etc. instead of one question about consumption of alcoholic beverages). Questions that are decomposed to a very detailed level may suffer from overestimation, whereas questions that are hardly decomposed may suffer from underestimation (Schwarz and Oyserman, 2001).

We used a decomposition strategy by first asking the frequency of behavior in a global category and then in a more specific category. This strategy was intended to separate the different enumeration strategies that respondents can follow. An example of the manipulated questions will illustrate this. Our target question was ‘How many hours and minutes did you spend on sports last week?’ Respondents need to retrieve two types of information to answer this question. Firstly, information about the number of days they had been engaged in sports last week. Secondly, the duration in hours and or minutes for each day they had been engaged in sports. Respondents may verbally express this enumeration, and while doing so mix up the two types of information (e.g., well, on Monday I went swimming for an hour, on Wednesday I exercised for twenty minutes, etc.). This enumeration is likely to be verbally expressed, and this increases the chance that respondents give a mismatch answer.

With the manipulation we tried to trigger separated retrieval of two types of information. Thus, for the easy version of the question, we first asked an extra question that directly asks retrieval of days, i.e., 'How many days did you spend on sports last week?' Then, the target question was asked. For the difficult version of the question, only the target question was asked. Thus, respondents answering the difficult question are not asked to verbally express their enumeration of days, whereas respondents answering the easy question are. In the questionnaire two questions (Q12, time spent on sports and Q14, time spent on walking) were manipulated in this way.

7.2.6 Operationalization 'ambiguous' versus 'non-ambiguous' questions

Hypothesis 4 concerns the difference between ambiguous and non-ambiguous questions. According to Churchill's (1978) 'procedural problem maxim', in conversations hearers will repair ambiguousness before answering questions. He distinguishes two types of 'procedural problems': missing information in the utterance of the speaker and information in the utterance of the speaker that the hearer disagrees with. The missing information problem can be caused by a lack of hearing, a lack of understanding the language, a lack of comprehending the meaning of the speaker's utterance, a lack of sufficient specificity in the speaker's utterance, or because the hearer does not know the answer. Churchill proposes that these five problems are hierarchically ordered. Here we are only interested in the lack of sufficient specificity as a cause of procedural problems, since this can be seen as a communication problem typical for survey interviews. Respondents usually understand the general meaning of a question, but may have problems with the way in which concepts are specified. For example, the question "how many days a week do you watch television?" may trigger questions like "does a week count five (week) days or seven days?", "does 'you' include my family or is it just me?", "does watching television mean following a program with full attention?", etc. (see Belson 1981)

In our definition of 'ambiguous' questions we face the problem that conversational questions are often ambiguous questions as well; they are more clear with respect to their general meaning, but not with respect to specificity. In ordinary conversations it is unusual to give exactly defined terms.

The 'question appraisal system' (Willis and Lessler 1999), gives several examples of problematic questions. Especially the following question is illustrative in this context:

(1) 'Do you have a car?'

In an ordinary conversation, this question would be informative enough, according to Grice's (1975) maxim of quantity. The words used in this question are probably frequently used. For example, the noun and verb are frequently used and easily retrievable from memory. The Dutch equivalents have a high frequency in the Corpus Spoken Dutch database (i.e., more than ten times the mean Corpus frequency) Therefore, we can define question (1) as a conversational question. However, as Willis and Lessler state, when the question is to be used

in a survey it is not clear, because it contains *undefined common terms*. Therefore, they suggest rephrasing the question as follows:

(2) 'Does anyone in your household now own or lease a car, truck or other type of vehicle?'

Question (2) would be considered as a non-ambiguous question. It can also be considered as a formal question; these types of utterances are unlikely to be expressed in ordinary conversations. Thus, question (1) and (2) cannot be compared correctly with respect to their ambiguity, as their conversational character is also different. To correctly compare ambiguous and non-ambiguous questions, next to a conversational, ambiguous question version ('Do you, or anyone at your home have a car?') and a formal non-ambiguous version ('Including all licensed vehicles, do you or anyone in your household own a car or motorcycle?'), we used a formal ambiguous version ('Do you, or anyone in your household own a car?'). However, the non-ambiguous version is a long question with a lot of specifications (i.e., 'including all licensed vehicles' 'you or anyone in the household', 'car or motorcycle'). The difference between ambiguous and non-ambiguous questions should not be too obvious. Respondents may get irritated about the number of specifications, and interviewers are, as a result of that irritation, likely to reword the question. Therefore we also tried to manipulate question wordings in a more subtle way. For example, we derived question manipulations from Mallison's findings (2002) on the question 'In general would you say your health is excellent, very good, good, fair or poor?'. She concludes from her study that (elderly) respondents have difficulties with this question. It appeared to be unclear to the respondents whether health should be compared to their peers (i.e., the elderly) or to the whole population or perhaps should be compared to their health when they were young. Different respondents appeared to use different frames of reference. Therefore we could use this question, originating from the Short-form General Health Survey (MOS-SF-20, König-Zahn et al. 1993) to manipulate ambiguity of the question. We formulated four different versions of this question, as depicted in Table 7-7.

Table 7-7 Four different wordings of the question concerning one's general perception of health

	Non-ambiguous	*Ambiguous*
Conversational	According to your age, do you have a very good, a good, reasonable or bad health?	Do you have a very good, a good, reasonable or bad health?
Formal	As compared to your peers, would you consider your physical health situation as very good, good, reasonable or bad?	Would you consider your health situation as very good, good, reasonable or bad health?

7.2.7 Construction of different questionnaire versions

With the different manipulations we obtained either two, three or four different versions of the same question. For example, the manipulation of ambiguity and conversational character of the question about one's general health perception (Q1) yielded four different versions (see

Table 7-7). The manipulation of wording of questions and alternatives for the assertions also yielded four versions. To enable between-respondent comparisons, these different versions were divided over different questionnaires. In each questionnaire one of the versions of a question was included.

Each questionnaire was divided into nine sections that dealt with the same topic. As far as possible, the same condition was used for all questions within a section. Finally, to control for question-order effects, the order of the nine sections was varied to some extent across the questionnaires. Of course it is not possible to include all 362,880 (the faculty of 9) possible orders of nine sections. All questionnaire versions ended with the same three sections concerning 'health contacts', 'body measures' and background questions. Some variations were made across the remaining six sections, which yielded six different questionnaire versions. In Appendix 7-2 an overview is given of the order of the manipulations and the separate sections within all six questionnaire versions.

Furthermore, in the ESS-study (chapter 6) we found that frequent changes in the answering format appeared to confuse respondents and urged interviewers to notify the respondent of the required answering format, even in case of a 'yes/no' format. This might be a serious disturbing factor, as it might be the continuous change in answering formats, instead of the specific manipulation of the answering format itself that causes respondents to give mismatch answers. Therefore, we had to be aware of the order of the questions with manipulated answer alternatives, and avoid continuous switches in answering formats. The response format is different for choice questions as compared to assertions. For choice questions either listed alternatives that are incorporated within the question or implicit alternatives are used. As discussed in section 7.2.3, for assertions, the listed alternatives are only presented in the introduction statement of the battery of assertions. For each assertion the respondents need to remember what these two or three alternatives were. We tried to use the choice questions as a buffer between two batteries of assertions. In this way we aimed to distract respondents from the alternatives used in a previous battery of assertions. Thus, the three batteries of assertions were alternated with choice questions. This ensured that a battery of assertions was hardly ever immediately followed by another battery of assertions with different response alternatives. Such adjacent batteries of assertions occurred only once in two questionnaire versions (i.e., questionnaire versions 5 and 6).

7.3 Procedures in conducting the interviews, response rate and coding

A CATI-program was written that took care of a user-friendly administration of the six different questionnaire versions and call-management. The latter comprised keeping count of the number of call-attempts (and deleting a telephone number from the list after six unsuccessful attempts), non-response cases, and schedules for appointments. The CATI-program also enabled digital audio recordings of the interviews (i.e., CARI, Computer Audio Recorded Interviewing).

The design of the screens displayed by the CATI-program (see Figure 7-1) fulfilled the requirements as mentioned by Hansen and Couper (2004, p. 344). For example, we used a

consistent screen design, and made sure that the questions, response alternatives, instructions and navigation buttons were immediately recognized as such.

Figure 7-1 Design of the CATI-Screen

To test the CATI-program, we used five of the basic testing procedures as listed by Tarnai and Moore (2004). The first procedure was 'question-by question testing', mainly to check the accuracy of the manipulations in question wording. Second, we used 'testing by task' to check branching of the questions (which in our case was fairly simple, involving two or at most four filter questions per questionnaire version). Testing by task also involved workability of warning messages in case of extreme scores or to prevent typing errors. Third, 'scenario testing' was done, mainly focused on the call-management features of the program and the random assignment of questionnaire versions to respondents. Fourth, 'data testing' was done, which involved not only a check of accuracy and completeness of the entered responses, but also a check of the quality of the digital sound files. And finally, a research assistant conducted 6 pretest interviews (one interview per version of the questionnaire) with actual respondents. This pre-testing also served as a final test of the questionnaire. The assistant checked whether it was possible to easily pronounce the questions, and whether the questionnaire appeared to respondents as a coherent and sensible survey. The latter was important, as the questionnaire in fact was a collection of questions from several different surveys. Respondents did not seem to notice this, as inferred from the absence of specific comments on the contents of the questionnaire.

The pre-testing interviews revealed that the phrase 'If I may ask…' that initially preceded questions concerning the respondents' height and weight (Q35 and Q36) actually stressed the sensitivity of the question, increasing the chances of item non-response. Therefore, this phrase, which was actually intended to decrease the sensitivity of the question, was deleted from the final question wording.

Although Tarnai and Moore (2004) in their study on effectiveness of testing methods conclude that testing by professional staff using simulated data was most effective and efficient, we did not apply this method. Simulation of survey data appears to be most effective in the detection of branching errors and response range errors. As our questionnaire did not include complex skip-patterns, and had a plain and easy response format for most questions, detection of branching and response range errors were not such important goals in pretesting. The feasibility and accurate manipulation of the question wordings were much more important concerns. Tarnai and Moore additionally found that a question-by-question review by interviewers better enabled detection of screen features and question wording errors. From our tests of the CATI-program we concluded that the program was working as intended, and could deal with unexpected situations.

7.3.1 Preparing procedures for the field work

The experiment was conducted in February and March 2004. In December 2003, a sample of telephone numbers was drawn from a website with telephone listings of households in all local communities in the Netherlands. A fairly complex procedure was used in order to obtain a stratified sample of households, according to the number of inhabitants of the local communities.

Candidate-interviewers were recruited by means of adds in several announcement facilities of the Faculty of Social Sciences (the educational website, the university newspaper, and bulletin boards in the university buildings). The candidates were selected by means of the following criteria: some experience and affinity with interviewing, social skills, and availability for all training and interview sessions. In order to prevent learning effects from earlier jobs as an interviewer, it was made sure that the interviewers did not have too much experience. The interviewers were all female social-science students aged between 19 and 28, who did not know each other. The interviewers were financially compensated, and received a certificate for their training hours when they had attended all training-sessions and had interviewed four evenings. Fortunately, all interviewers were able to interview for four nights, exactly according to the original schedule.

7.3.2 Interviewer training and fieldwork

The interviewers were split up into four interviewer groups (of three interviewers each), which were separately trained within one week before their first interview evening. Each training started with two sessions of approximately six hours of basic interviewing techniques each. During these two sessions, interviewing techniques were discussed and practiced by means of role-playing. On the third instruction day each of the three interviewers within the session had to conduct a test interview with a respondent from the sample, which was recorded. All interviewers participating in the session listened to each other's test interviews, and problems occurring during the interviews were discussed. Finally, the interviewers interviewed for four evenings, scheduled within a period of two weeks.

During the fieldwork, the interviewers were monitored using the digital recordings. All interviews of the first evening of each interviewer were listened to, and discussed with her at the second evening. For the other evenings parts of a few interviews were listened to, only for the interviewers who appeared to behave problematically on their first evening. If necessary, interviewers were instructed to improve their behavior. This was especially necessary for some interviewers with regard to neutral probing, adequate clarification of questions, and persuading reluctant respondents.

Each interview evening took place between 6.00 p.m. and 9.30 p.m. Interviewers were instructed to call numbers provided by the CATI-program and to interview the first person available in the household. We did not use random respondent selection within the household (e.g., selecting the household member whom had last had his birthday) but the first person who answered the phone.

In order to avoid confounding of interviewers and questionnaire versions, it was important that all interviewers administered an equal amount of all versions. We also had to assure that the number of different questionnaire versions was distributed evenly over all interview evenings. Therefore, each evening different questionnaire versions were systematically assigned to the interviewers. The interviewers all started with a different questionnaire version, and were assigned all versions, in different orders. For each interviewer the same questionnaire version was used for five respondents in a row. For the next series of five respondents another questionnaire version was assigned to the interviewer. As the interviewers on an average interview evening each conducted about twelve interviews, each interviewer was confronted with about three different versions of the questionnaire in the same evening. Fortunately the shifts in versions of questionnaires (and hence shifts in question wording) had no consequences for the way interviewers read the questions. Questions were read exactly as worded in nearly 95% of the cases.

Although interviewers knew that different question wordings were used, they were not informed about the actual hypotheses that were to be tested in the study. They were told that different question wordings were used to control for question wording effects on response distributions. To further distract them from speculations about the true goals of the study, the interviewers were told that the primary goal of the study was to examine relations between solidarity in health issues and the respondent's (perceived) health status. As examples, fake research questions were presented to interviewers, such as 'what are the differences between respondents with unhealthy behaviors and respondents with healthy behaviors with respect to their opinion on public health issues?' The interviewers were not informed about the details of the number of different questionnaire versions and the order of assignments of questionnaires to respondents. After their last interview session the interviewers were informed about the origin of the types of questions, and the actual goals of the study. This came as a surprise to them, and therefore we assume that interviewers were not suspicious about the goals of the study, and did not adapt their behavior to unhoped-for expectations.

7.3.3 Response rates

In Table 7-8 the response rate of all telephone calls is shown. We used the AAPOR Standard Definitions (Third edition, AAPOR 2004) in order to define our case codes and outcome rates. In total 1525 different telephone numbers were dialed, and eventually 40% of these calls resulted in a completed interview (based upon including both eligible and non-eligible cases). The largest number of non-response calls are refusals, as could be expected. In 34.4% of the calls respondents could not be persuaded to engage into an interview. In this study the number of non-eligible cases is high, due to a large number of non-working numbers (code 4.30 in Table 7-8). This might be caused by the fact that the telephone numbers were sampled about two months before data collection started. The 'Response Rate 1' (or minimum response rate) excludes the non-eligible cases, and is 44.5% (610 / (610 + 524 + 236) * 100%).

Table 7-8 Response of all numbers called

	N	Percentage
Complete interview (1.0)	610	40.0%
Eligible, Non-interview (2.0)	524	34.4%
Refusal (2.11)	500	32.8%
Break-off (2.12)	13	0.9%
Other (2.30)	2	0.1%
Physically or mentally unable/incompetent (2.32)	1	0.1%
Language problem (2.33)	8	0.5%
Unknown Eligibility, Non-Interview (3.0)	236	15.5%
Not attempted (3.11)	174	11.4%
Never answered (3.13)	57	3.7%
Other (3.90)	5	0.3%
Non-Eligible (4.0)	155	10.2%
Fax/data line (4.20)	5	0.3%
Business, government office, other organization (4.52)	12	0.8%
Non-working and disconnected number (4.30)	138	9.0%
Total	1525	100.0%

Due to a high number of telephone numbers that could not be attempted for six trials (code 3.11), the number of cases with unknown eligibility is fairly high. Telephone numbers that could not be contacted by interviewers within their scheduled interview evenings, could not be shifted towards other interviewers, and had to be recorded as non-contacts before the minimum requirement of six attempts was reached. Therefore the Cooperation rate ('COOP 1' or 'Minimum Cooperation Rate 1', AAPOR, 2004) in this case is also relevant. This is the proportion of interviews of all eligible units ever contacted, and is 53.8% (610/(610+524) * 100%).

As compared to Dutch surveys, our response rate is not low. As De Leeuw and Heer (2002) note, The Netherlands has a high refusal rate and the response rates have been

declining over the years, at least until 1997. No general information of response rates since 1997 is available, but individual recent surveys also indicate that a response rate for a Dutch telephone survey will not easily turn out to be above 50%. For example, a more recent survey concerning 16-minute telephone interviews about online newspaper reading (conducted in December 2002) yielded a response rate of 41% (De Waal, Schonbach and Lauf 2004).

The interviewers did not significantly differ with respect to the response rates they achieved. They did also not differ with respect to the percentage of female and male respondents and the average age of the respondents they interviewed. However, they did differ with respect to the number of interviews they conducted (i.e., between 42 and 60 interviews per interviewer). The questionnaire versions were equally distributed over interviewers; they all had conducted at least five interviews with each version. There were also no significant differences between the six questionnaire versions for the respondent variables[13] age, education and gender. We can therefore conclude that the random assignment of the six questionnaire versions was successful.

7.3.4 Coding of the data

The number of completed interviews (610) comprised a dataset of 25,670 Q-A sequences. The digital recordings of all 610 completed interviews were of good quality. The sound files were transcribed and coded by three transcribers and three coders (all graduate students) and for a small part by the researcher. The three coders initially transcribed and coded a few interviews, to get acquainted with the whole process (e.g., to learn that for coding sometimes listening to original sound files is required). Later on, the coders only coded interviews that were transcribed by the others, who were not trained to use the coding scheme. In this way, we had an 'assembly line' of production (three coders being specialized in coding, three transcribers selected for their rate and quality of typing). To make sure coders were randomly assigned to interviewers, respondents and questionnaire versions, all interviews were randomly distributed to the coders. Nevertheless, due to an uneven availability of the coders, it was not possible to assign an equal amount of interviews to each coder. Two coders each coded about 40% (i.e., together 81%) of the interviews, the third coder coded 18% of the interviews and the remaining 1% was done by the researcher. The coders were instructed by means of a coding manual. The first several interviews they coded were evaluated by the researcher, and discussed individually.

It took about three months to transcribe and code all 25,670 Q-A sequences, which comprised transcription of more than 85 hours of speech, and assignment of codes to 136,619 utterances. The complete dataset was checked for the occurrence of rare codes and typical

[13] We compared the distribution of some respondent characteristics of our sample with figures concerning the Dutch population from the Dutch Statistics website (on January 1st 2004). From this comparison it turns out that the proportion of female respondents (59%), is slightly higher than this proportion in the Dutch population (50.5%). The proportion of respondents older than 65 is also slightly higher (18%) than in the Dutch population (14%). Finally, for persons between 15 and 64 years old, the proportion of respondents with higher education is higher (33%) as compared to the Dutch population (23%), whereas the proportion of respondents with at most primary education (6%) is lower than in the Dutch population of 15-64 year olds (12.5%).

errors, with search options of the Sequence Viewer program (Dijkstra 2002). For example, questions from respondents (code 'RQ0') or answers from interviewers (code 'IA0') are unlikely to occur, and therefore all instances of such codes were adapted if they were incorrect.

In order to assess the reliability of the coding, the researcher coded a random sample of 10% of all Q-A sequences, excluding the Q-A sequences originally coded by the researcher. These codes were compared with the original coding. The percentage of agreement in the two coded files appeared to be 82%, and the Cohen's Kappa Value was .81. According to the scale that Landis and Koch (1977) proposed to describe the degree of agreement, a Kappa Value of .81 can be considered as "almost perfect". However, this agreement is based upon the complete code string. For example, an 'adequate answer' (code 'RA0AA') and an 'invalid answer' (code 'RA0AI') are considered as different, despite the fact that they are coded the same for four of the five variables. The reliability may also be assessed for the individual variables coded. The weighted Kappa in the Sequence Viewer Program uses the number of code variables that are coded differently to weight the disagreement between the two codes. In this way, the weighted kappa takes into account that coders may have assigned partly the same code to an utterance. The weighted Kappa Value was .90. Thus, the reliability of the coding was rather high, also as compared to values obtained with other coding schemes (cf. Table 3-10 in section 3.5.4)

For the analyses in this experiment, the most important aspect of the quality of the coding is the recognition of mismatch answers. For a correct test of hypotheses it is important that the recognition of occurrence of mismatch answers is independent from the coders. The percentage of agreement for assigning the codes 'adequate answer' and 'mismatch answer' appeared to be 96% (Kappa and weighted Kappa = .87). It turned out that, taking the reliability coding of the researcher as a 'gold standard', the three coders failed to recognize 7% of the mismatch answers, whereas 9% of the mismatch answers were falsely recognized. From these figures we may conclude that the recognition of mismatch answers had good inter-coder reliability.

7.3.5 Operationalization of occurrence of mismatch answers

To test the hypotheses, we will compare the percentage of Q-A sequences in which a mismatch answer occurs for all questions. When the percentage of Q-A sequences with a mismatch answer is higher for a conversational question than for a formal question, we conclude that this difference with respect to question wording caused this difference. It is important to eliminate other factors that might have caused differences. If questions are not read as worded, this conclusion may not be valid. In that case we do not know if the chance of a mismatch answer is caused by the intended manipulation. Furthermore, it is important to distinguish between immediately evoked mismatch answers and mismatch answers in a later stadium. For example, a mismatch answer may be preceded by a request for clarification. The interviewer action (e.g., repeating the question or clarifying the question) that is likely to occur after such a request might evoke a mismatch answer rather than the original question.

Finally, the mismatch answer may be an initial response that is subsequently self-corrected by the respondent (i.e., without interviewer probing). It is important to distinguish such mismatch answers from mismatch answers that are corrected by means of interviewer probing. Below we will give descriptions of different situations that should be taken into account in distinguishing mismatch answers that may be caused by factors different from our manipulations.

Initial question reading

Question reading is considered 'as worded' when an interviewer reads *all* components exactly as worded or only with minor deviations (i.e., false starts, 'uhs' etc.). We assume that such minor deviations do not disturb the manipulation of question wording significantly. Q-A sequences are excluded from the analyses if they do not meet this criterion. Q-A sequences are included when respondents interrupt the introductory part of the question with requests for repetition, neutral acknowledgements or comments, whenever the interviewer proceeds reading the question as worded, or just repeats the question exactly as worded. However, whenever an interviewer replies to this request in a problematic way (i.e., not repeating the question exactly as worded) the Q-A sequence is excluded from the analyses.

A different issue is interruption of the question proper. We structured most questions in such a way that the question component is presented last (see section 7.2.3). This component is the most important part of the question, without which it is hardly possible to infer the meaning of the question. Although, because of this principle, this structure is helpful to decrease the chance of interruptions, they may of course occur nonetheless. Thus, when respondents interrupt question reading with an answer, due to this interruption, the question did not correspond to the manipulation of the question, and hence the question should be viewed as significantly changed. Such Q-A sequences are excluded from the analyses.

We distinguish two different deviations of initial question reading that are considered non-problematic, i.e., fulfilling the criterion to be included in the analyses.

a. Neutral addition of phrases to introduction wording (i.e., the interviewer adds a phrase like 'the first assertion is' or 'the next assertion is').
b. Neutral deletion or rewording of introduction wording (i.e., the interviewer does not read 'Now I will read some general questions' or deletes/changes words in the introduction).

We distinguish seven different deviations of initial question reading that do *not* fulfill the criterion to be included in the analyses:

c. Deletion or serious rewording of definition (i.e., the interviewer does not explain what is meant by sports, weekdays etc.).
d. Deletion of alternatives (i.e., the interviewer does not read any or only some of the alternatives).

e. The interviewer rewords the question, changing the manipulation of the question, or even the meaning of the question.
f. The interviewer rewords the question in a suggestive way.
g. The respondent interrupts question reading, giving an invalid answer.
h. The respondent interrupts question reading with a request for clarification.
i. The interviewer incorrectly skips the question (i.e., fills in answer without verification).

As Table 7-9 shows, in over 94% of the 25,670 Q-A sequences the question is initially read exactly as worded, including all scripted introductions and definitions. The occurrence of deviations from question reading was not related to specific types of questions.

Table 7-9 Frequency of different types of question reading

	Frequency	Percentage	Cumulative percentage
All parts of question read exactly as worded	24219	94.35%	94.35%
a) Neutral addition to introduction	701	2.73%	97.08%
b) Neutral deletion of (words in) introduction	503	1.96%	99.04%
c) Deletion/rewording definition	39	0.15%	99.19%
d) Deletion of answer alternatives	0	0.00%	99.19%
e) Rewording of question	40	0.16%	99.35%
f) Suggestive question	19	0.07%	99.42%
g) Respondent interrupts with invalid answer	129	0.50%	99.92%
h) Respondent requests clarification before question	10	0.04%	99.96%
i) Interviewer incorrectly skips question	10	0.04%	100.00%
Total	25670	100.00%	100.00%

Interactions preceding mismatch answers

Interactions that occur between initial question reading and the initial answer create another interpretation problem for the relation between question wording and the occurrence of mismatch answers. Although the question may initially be read as worded, the fact that something else occurs before the respondent gives an answer troubles conclusions that the chance of a subsequent mismatch answer or adequate answer was caused by question wording only. Especially when the respondent asks for clarification of the question and the interviewer subsequently gives a clarification in her own wording rather than repeating the question, we are faced with a problem of interpreting the cause of the type of (mismatch or adequate) answer. Therefore, all Q-A sequences in which, after initial question reading but before the respondent's first answer, the interviewer did something else than exactly repeating the question were excluded from the analysis. Q-A sequences with adequate answers that were changed into mismatch answers that indicated nearly the same answer, and Q-A sequences during which no direct answer was given at all were also excluded from the analysis. These deletions based upon the interactions preceding mismatch answers comprised 981 Q-A sequences. Deletion based upon initial question reading comprised 247 Q-A

sequences. In this way, 1228 of the original 25,670 Q-A sequences were excluded from the analyses, and thus 24,442 eligible Q-A sequences were available for analysis.

Mismatch answers and invalid answers that were self-corrected (i.e., the respondent changed the mismatch or invalid answer into an adequate one, without intervention of interviewer probing) were considered as adequate answers. According to these definitions, mismatch answers occurred in 19% of the remaining 24,442 Q-A sequences.

7.4 Results

In this section we will describe the results of the analyses in relation to the hypotheses as formulated in section 7.1.1. As all our hypotheses are formulated in one direction, they will be tested with one-tailed significance tests, using a significance level of $p < 0.05$. The results for hypothesis 1 and 2 (concerning the effects of conversational and formal wording of questions and answer alternatives) will first be described for the assertions. After discussion of the effects of questions and alternatives (section 7.4.1), an overall analysis is presented in section 7.4.2. This analysis takes effects of both question wording and types of alternatives into account, and other variables that are related to the chance of mismatch answers (such as the respondents' age). The effects of question wording and types of alternatives for choice questions (hypothesis 1, 2a, and 2b) will be described in sections 7.4.3, 7.4.4 and 7.4.5 respectively. Finally, in sections 7.4.6 and 7.4.7 the effects of difficulty (hypothesis 3) and ambiguity (hypothesis 4) will be described.

7.4.1 Effects of formal and conversational assertions and alternatives (H1 and H2a, assertions)

Hypothesis 1 concerned the difference in the occurrence of mismatch answers for conversational and formal question wording, whereas hypothesis 2a concerned the difference in the occurrence of mismatch answers with respect to the wording of alternatives. Table 7-10 shows the percentage of Q-A sequences with and without mismatch answers, for the wording of the assertions of the three different batteries of assertions (i.e., testing hypothesis 1). These results do not show a confirmation of this hypothesis for all assertions.

Table 7-10 Percentage of Q-A sequences with mismatch answers for all assertions

GHPQ assertions (n = 8)					
Q-A sequences	Conversational		Formal		Total
With mismatch	379	**11%**	153	**11%**	532
Without mismatch	3108	89%	1255	89%	4273
Total	3397	100%	1408	100%	4805
χ^2= n.s. (1.5% Q-A sequences excluded)					
Public Health assertions (n = 6)					
Q-A sequences	Conversational		Formal		Total
With mismatch	466	**19%**	191	**17%**	657
Without mismatch	1967	81%	906	83%	2873
Total	2433	100%	1097	100%	3530
χ^2= n.s. (3.6% Q-A sequences excluded)					
Government and Health assertions (n = 3)					
	Conversational		Formal		Total
Q-A sequences					
With mismatch	183	**22%**	151	**16%**	334
Without mismatch	647	78%	814	84%	1461
Total	830	100%	965	100%	1795
χ^2= 12.07, $p < 0.01$ (1.9 % Q-A sequences excluded)					

For the eight GHPQ-assertions, the percentage of Q-A sequences with a mismatch answers was 11 percent in both conditions. Although for the six 'Public Health' assertions, the percentages of Q-A sequences with a mismatch answer were in the expected direction (19% for the conversational versions of the questions; 17% for the formal versions) the difference was not significant.

For the 'Government and Health' assertions we did find a significant overall difference between the conversational and formal versions. As the results in the table show, the conversational versions yield mismatch answers in 22% of the Q-A sequences, whereas for the formal versions this percentage is 16%. Table 7-11 shows the percentage of Q-A sequences with and without mismatch answers for the wording of the alternatives (i.e., testing hypothesis 2a). These results show a confirmation of this hypothesis in all cases: assertions with formal answer alternatives yield more mismatch answers than assertions with conversational alternatives.

Table 7-11 Percentage of Q-A sequences with mismatch answers for alternatives of all assertions

GHPQ assertions					
	Alternatives				
Q-A sequences	Conversational 'yes', 'maybe' or 'no'		Formal 'true', 'possibly true' or 'false'		Total
With mismatch	86	**4%**	446	**19%**	532
Without mismatch	2303	96%	1970	82%	4273
Total	2389	100%	2416	100%	4805
χ^2= 269.42, p < 0.01 (1.5% Q-A sequences excluded)					
Public Health assertions					
	Alternatives				
Q-A sequences	Conversational 'yes' or 'no'		Formal 'agree' or 'disagree'		Total
With mismatch	173	**10%**	484	**27%**	657
Without mismatch	1589	90%	1284	73%	2873
Total	1762	100%	1768	100%	3530
χ^2= 179.59**, p < 0.01 (3.6% Q-A sequences excluded)					
Government and Health assertions					
	Alternatives				
Q-A sequences	Conversational 'yes' or 'no'		Formal 'strongly agree', 'agree', 'neutral', 'disagree' or 'strongly disagree'		Total
With mismatch	91	**10%**	243	**27%**	334
Without mismatch	813	90%	648	73%	1461
Total	904	100%	891	100%	1795
χ^2= 87.72** p < 0.01 (1.9 % Q-A sequences excluded)					

For the eight GHPQ-assertions, the difference in the percentage of Q-A sequences with a mismatch answer is 19% (formal alternatives) versus 4% (conversational alternatives). For 'public health' assertions the difference in the percentage of Q-A sequences with a mismatch answer is 27% (formal alternatives) versus 10% (conversational alternatives). For the response alternatives of the 'Government and Health' assertions, the difference in the percentage of Q-A sequences with a mismatch answer is 27% (formal alternatives) versus 10% (conversational alternatives).

In the latter case, not only the wording of the alternatives was manipulated, but also the number of alternatives (two alternatives in case of conversational wording, and five alternatives in case of formal wording). This was done in order to compare common survey practice (the five alternatives) with alternatives that in our view are best suited for assertions, i.e., minimizing the chance of mismatches because of the conversational character of responding with just 'yes' or 'no' (see also section 7.2.4). We expected to find even more pronounced differences between both types of alternatives than for the more properly manipulated alternatives of both other types of assertions. Although this does not appear to be the case, the differences for all types of assertions are quite striking.

Assertions that concern opinions (i.e., the 'Public Health' and the 'Government and Health' assertions) yield more mismatch answers than assertions that concern perceptions of respondents' own health (the GHPQ-assertions). In another study (Draisma, Dijkstra and

Ongena forthcoming), it was shown that respondents provide more verbal considerations when they have to answer assertions that concern their opinion than when they have to answer perception assertions. This is especially the case when they are asked about their opinion on social controversial topics such as the costs of public health. They are more likely to explain and justify their answer. In doing so they might be less focused on the appropriate response categories, and in their explanations produce more mismatch answers.

These results clearly confirm our hypothesis 2a; conversational alternatives yield less mismatch answers than formal alternatives.

The results do not clearly confirm hypothesis 1. Although a significant effect for the 'Government and Health' assertions was found, further analysis showed that this difference was not significant if we compared both versions of the question wording within each condition of both versions of the alternatives (see Table 7-12). This difference is caused by an uneven distribution of the six ballots over the four conditions (i.e., the 2*2) design for question wording and types of alternatives). Although the ballots were equally distributed over the assertions, within the assertion conditions we had an uneven number of ballots (i.e., each three), which had to be divided over two conditions (i.e., formal and conversational alternatives). Thus, the manipulations of question wording and types of alternatives are not independent, and tests of effects should consider both manipulations instead of one. Such an analysis is presented in the next section.

Table 7-12 Number of Q-A sequences with a mismatch answer in a 2*2 design

Government and Health assertions						
	Alternatives					
Assertions	Conversational		Formal		Total	
Conversational	31	11%	152	27%	183	22%
Formal	60	10%	91	27%	151	16%
	91	10%	243	27%	334	19%
Public Health assertions						
	Alternatives					
Assertions	Conversational		Formal		Total	
Conversational	144	12%	322	27%	466	19%
Formal	29	6%	162	28%	191	17%
	173	10%	484	27%	657	19%

A similar analysis of the 'public health' assertions showed a significant question wording effect within the conversational alternatives, but not within formal alternatives. When conversational alternatives were used with conversational assertions they yielded mismatch answers in 12% of the cases, whereas with formal assertions they yielded mismatch answers in only 6% of the cases ($\chi 2 = 13.96$, $p < 0.01$).

In fact this difference may indicate that the conversational assertions not only stimulate respondents to give conversational answers (and thus adequate answers when the response alternatives are conversational), but also stimulate respondents to give less precise answers or to give no direct answer at all. It might be possible that conversational questions increase the

likelihood that respondents give no answer at all. In our computation of the percentage of mismatch answers, we excluded Q-A sequences during which respondents did not give any direct answer, or gave an answer only after the interviewer's probe (see section 7.3.5). Analysis of these excluded Q-A sequences may reveal whether conversational questions indeed generated more non-answers than formal questions. However, the number of Q-A sequences with no substantial answer (i.e., n = 40 for all assertions) is too low to be able to compare differences between conversational and formal questions.

7.4.2 Modeling effects of questions, respondent and interviewer variables (H1 and H2a, assertions)

Although hypothesis 2a was confirmed by univariate analyses, it appeared that the manipulations of the question wording and alternative type effects were not independent from each other. Therefore we should consider effects of both manipulations in an overall analysis. Respondent characteristics such as age and level of education, are also related to the chance of mismatch answers, and thus need to be taken into account as well. For example, older and lower educated respondents are more likely to give mismatch answers than younger and lower educated respondents.

Our dependent variable, the occurrence or absence of a mismatch answer in a Q-A sequence, is a dichotomous variable. Therefore we used a logistic regression analysis to model the effect of all variables. In the three models presented in Table 7-13, we subsequently take the effects of both question wording, types of alternatives, and respondent variables into account. The unstandardized regression coefficients and the odds ratio (exponent of B, Exp (B) in the table) of significant variables are reported. For the categorical variables question wording, type of alternatives and level of education of respondents, we used the following values as a reference category: the 'conversational questions', 'conversational alternatives' and 'higher education'.

Table 7-13 Logistic regressions for the odds of mismatch answers occurring in a Q-A sequence

	Model 1		Model 2		Model 3	
	B	Exp (B)	B	Exp (B)	B	Exp (B)
Question variables:						
Question wording Formal	- 0.33	0.97	- 0.21	0.81	-0.19	0.82
Alternatives Formal	1.39**	4.03	0.32**	3.72	0.39**	4.00
Q * A (Formal*formal)			0.24	1.27		1.20
Constant	-2.58**	0.08	- 2.53**	0.08	-4.30**	0.01
Respondent variables:						
Age of respondent					0.03**	1.03
Education of respondent (ref: higher education)					0.58**	1.80
No education/primary					0.51**	1.66
Lower vocational					0.42**	1.52
Vocational education					0.08	1.09
Pre-university education						
Akaike Information Criterion	8037.54		8032.40		7562.38	

Model 1, including only question variables, summarizes the results that were presented in section 7.4.1, but in this case both types of manipulations are included in one analysis. The odds of a mismatch answer in a Q-A sequence with formal alternatives are four times higher (odds ratio = 4.03) than those of a Q-A sequence with conversational alternatives. The effect of question wording is not significant.

In model 2 the interaction of question wording and types of alternatives is taken into account. The odds of a mismatch answer in a Q-A sequence with formal alternatives are still almost four times higher than those of a Q-A sequence with conversational alternatives, but the effects of question wording, and the interaction between question wording and wording of the alternatives are not significant.

The parameters appear to remain almost the same when both respondent variables are included (model 3). The effects of the respondents' age show that for each year a respondent is older, the odds of a mismatch answer occurring, increases with 3%. The lower levels of education of respondents also differ significantly as compared to the highest levels of education, indicating that the odds of a mismatch answer are higher for respondents in the three lowest levels of education than for respondents in the two highest levels of education (i.e., pre-university and higher educated respondents). However, the effects of respondent variables did not substantially change the effects we found for wording of the alternatives. No significant interaction effects between respondent and question variables could be found, and therefore are not presented here. We found that model 3 predicted the odds of mismatch answers more accurately (as indicated by a lower value for the Akaike Information Criterion, that is based upon the log likelihood for the model, but adjusts for the number of parameters). Model 3 does not change our conclusion about the effects of wording of alternatives on the odds of a mismatch answer. However, in these analyses the nesting structure of the data

(respondents answered multiple questions of the same version in a battery) is not taken into account. Thus, our observations of mismatch answers occurring in Q-A sequences are not independent from each other, which is an assumption of ordinary logistic regression. In Table 7-14 an analysis is presented that takes this nesting structure into account.

Table 7-14 Generalized linear mixed models (PQL) for the odds of mismatch answers occurring in a Q-A sequence

	Model 4		Model 5	
	B	Exp (B)	B	Exp (B)
Question variables:				
Question wording Formal	-0.23*	0.79		
Alternatives Formal	1.54**	4.66	1.55**	4.71
Q * A (Formal*formal)	0.21	1.23	0.23	1.26
Constant	-3.02	0.05	-4.98**	0.01
Respondent variables:				
Age of respondent			0.03**	1.03
Education of respondent (ref: higher education)				
No education/primary			0.66**	1.93
Lower vocational			0.51**	1.66
Vocational education			0.37*	1.45
Pre-university education			-0.07	0.93
Random effects				
Random intercept	1.71	5.53	1.56	4.78
Akaike Information Criterion	54279.71[1]		54055.42	

[1]In the R-program, scaling of the log likelihood is different for generalized linear models as compared to ordinary logistic regressions. The Akaike Information Criterion of models 3 and 4 can only be compared with each other and cannot be compared to the Akaike Information Criterion from the ordinary logistic regression (i.e., model 1 and 2).

* $p < 0.05$, ** $p < 0.01$, n = 10130 Q-A sequences, 610 respondents,

Model 4 is a generalized linear mixed model and includes only question variables. Parameter estimation in generalized linear mixed models is complicated because some kind of approximation is involved (Snijders and Bosker 1999). We used penalized quasi-likelihood (PQL) for which the approximation is around an estimate for the fixed and random part. An alternative approximation method is marginalized quasi-likelihood (MQL) for which approximation is around the estimated fixed part only. This latter method tends to underestimate the parameters. We tried to improve the fit of the model by including the respondent's age. As model 5 shows, age included as a continuous variable yielded a significant effect, indicating that the odds of occurrence of a mismatch answer increases with 3% each year a respondent is older. The lower levels of education of respondents again differ significantly as compared to the highest level of education. The odds of a mismatch answer are higher for respondents in the three lowest levels of education than for respondents in the highest level of education. The effects of respondent variables did not substantially change the effects we found for wording of the alternatives.

Models 4 and 5 cannot be compared directly with models 2 and 3. Small differences in parameter estimates and p-values may be due to a difference in parameter estimation, but also of taking the nesting structure into account.[14]

From the analyses we conclude that hypothesis 2a can be confirmed, the effects of types of alternatives are much stronger than the effects of question wording, and there is no interaction between question wording and the types of alternatives.

7.4.3 Effects of the conversational and formal wording of choice questions (H1, choice questions)

In Table 7-15, the percentages of Q-A sequences with mismatch answers are presented for 21 behavioral, factual and perceptual choice questions that were manipulated for question wording. We compared versions of questions for which the other manipulations (types of alternatives, difficulty and ambiguity of questions) were held constant. The last column of this table shows whether the difference in the percentage of mismatch answers was according to our hypothesis (+), opposite to our hypothesis (-) or no difference in the percentage of mismatch answers was found (0). It appears that for nine questions, the difference confirms our hypothesis (of which five are statistically significant), for twelve questions the difference is opposite to our hypothesis , and for three questions no difference was found. Overall, the results do not clearly confirm our hypotheses.

[14] We used the MLwiN program to estimate models with a random slope. From these analyses it appears that the parameter estimates for the effects of the wording of alternatives are similar to estimates without random slopes. However, the inclusions of a random slope in a model with explanatory variables at the respondent level also created convergence problems in the MLwiN program. We therefore assume that these estimations are not reliable. According to Snijders and Bosker (1999, p. 122), MQL and PQL approximations for models with a random slope cannot be used in reliable deviances tests in MLwiN. Moreover, testing random slopes is still a matter of active research.

Table 7-15 Difference in percentage of mismatch answers for choice questions

	Question wording		N	% missing[a]	χ^2	exp
Question	Conversational	Formal				
Q1 Perception of health	11%	10%	575	5.7%	n.s.	+
Q10 Last 12 months sports	2%	9%	565	7.4%	n.s.	-
Q10 Last 12 months sports (corrected for question order)	2%	2%	546	10.5%	n.s.	0
Q11 Days occupied with sports	22%	37%	163[b]	12.4%	n.s.	-
Q12 Hours/minutes sports (non-ambiguous, easy version)	8%	19%	142[c]	8.1%	n.s.	-
Q13 Ever walk 10 minutes	5%	12%	359[d]	7.0%	n.s.	-
Q13 Ever walk 10 minutes (corrected for question order)	5%	6%	349	7.2%	n.s.	-
Q14 Days occupied with walking	44%	36%	503[e]	7.0%	n.s.	+
Q16 Days watching Television (implicit alternatives only)	61%	50%	290	6.5%	2.67*	+
Q17 Time watching television	25%	25%	532[f]	9.0%	n.s.	0
Q21 Days using breakfast (implicit alternatives only)	44%	52%	337	12.7%	n.s.	-
Q22 Meat at dinner	30%	37%	337	6.6%		-
Q23 Fruit	34%	36%	556	8.9%	n.s.	-
Q24 Non-alcoholic beverages (implicit alternatives only)	38%	28%	316	18.1%	3.64*	+
Q26 Alcoholic beverages	31%	29%	364[g]	15.2%	n.s.	+
Q35 Body length	2%	3%	594	2.6%	n.s.	-
Q36 Body weight	6%	5%	604	1.0%	n.s.	+
Q37 Attitude towards body weight (conversational alternatives only)	22%	16%	400	2.2%	2.89*	+
Q39 Employment (yes/no)	2%	2%	599	1.8%	n.s.	0
Q40 Level of education	20%	22%	576	5.6%	n.s.	-
Q41 Age/Year of birth	3%	0%	602	1.3%	8.27**	+
Q42 Persons in household	15%	8%	596	2.3%	6.06*	+
Q43 Ownership of car (ambiguous versions only)	6%	13%	403	0.2%	n.s.	-

[a] Percentage of Q-A sequences that had to be excluded from all eligible Q-A sequences
[b] This question was only asked to respondents who had been engaged in sports the last 12 months
[c] This question was only asked to respondents in the condition 'easy question' and who had been engaged in sports the last 12 months
[d] This question was only asked to respondents in the condition 'easy question' and who ever walk 10 minutes on end
[e] This question was only asked to respondents who ever walk 10 minutes on end
[f] This question was only asked to respondents who watch television at least once a week
[g] This question was only asked to respondents who had drunk alcoholic drinks the last 12 months
*** p < 0.01, * p < 0.05 (according to one-tailed tests)*

Hypothesis 1 can be confirmed for the questions regarding the number of hours and minutes respondents watch television (Q16), the number of non-alcoholic beverages consumed (Q24), the respondent's attitude towards their body weight (Q37), the age of the respondent (Q41), and the number of persons in the household (Q42).

For quite a lot of questions it appeared that, contrary to our expectations, the formal version of the question yielded more mismatch answers than the conversational version. This especially holds for the questions Q10, Q11, Q12, and Q13. The high item non-response rate for Q11 and Q12 is caused by the fact that these questions were not asked when respondents indicated in the preceding filter question that they were not engaged in sports during the past twelve months. Older respondents are less likely to have been engaged in sports and more likely to give mismatch answers (as we already showed in section 7.4.2 for the assertions). Therefore, paradoxically, the respondents who should have benefited most from 'easy' worded questions were most often excluded from this analysis. Furthermore, Q11 was not asked in half of the versions of the questionnaire (because this question was used to manipulate difficulty, see section 7.4.6), and for Q12 only the non-ambiguous and easy versions could be compared for conversational character of the question wording.

Regarding questions 10 and 13, an order effect appeared to disturb our results. This effect was caused by the fact that these simple yes-no questions were in some questionnaire versions preceded by assertions with formal alternatives (i.e., 'true' and 'not true'). After such a battery of assertions, respondents were accustomed to use these formal alternatives, and therefore continued to use 'true' and 'not true' instead of the prescribed alternatives 'yes' and 'no'. Hence, such answers were coded as mismatch answers. If we discard Q-A sequences for which respondents answered these both questions with 'true' or 'not true', the difference between the formal and conversational versions virtually disappeared (see Table 7-15, Q10 and Q12, 'corrected for question order'). Although such a question order effect could also have contributed to a confirmation of our hypotheses we could not find evidence for such effects. For example questions with formal alternatives that were preceded by questions with conversational alternatives did not yield more mismatch answers than the same alternatives that were not preceded by conversational alternatives.

To summarize, for five of the 23 questions we found significant differences in the expected direction. These questions concerned Q16 (watching television), Q24 (non-alcoholic beverages), Q37 (attitude body weight), Q41 (age) and Q42 (persons in household). In four cases we found fairly strong differences the other way around. The results for all these four 'negative' questions however appeared to be disturbed by other factors, i.e., for two questions an order effect because of the format of the immediately preceding questions, and for other two questions a selection bias, causing that respondents who usually give most mismatch answers (older respondents) did generally not answer these questions.

7.4.4 Effects of conversational and formal alternatives for choice questions (H2a choice questions)

Hypothesis 2a concerned the difference in the occurrence of mismatch answers for conversational and formal wording of alternatives. As is shown in Table 7-16, hypothesis 2a can be confirmed for questions 16 (watching television) and 21 (use of breakfast). As expected, the conversational alternatives (vague quantifiers), yield less mismatch answers than the formal alternatives (numbers). For questions 15 (pace of walking) and 37 (attitude

towards body weight) no significant differences in the percentage of mismatch answers for conversational and formal alternatives were found. In general, the results give the impression that hypothesis 2a can be confirmed.

Table 7-16 Percentage of mismatch answers for the types of alternatives (conversational versus formal) for choice questions

	Alternatives		N	% missing[a]	χ^2	exp
Question	Conversational	Formal				
Q15 Pace of walking	26%	23%	536[b]	1.8%	n.s.	-
Q16 Days watching television (conversational questions only)	29%	40%	286	4.7%	3.83*	+
Q21 Days using breakfast (formal questions only)	44%	62%	192	14.3%	6.01*	+
Q37 Attitude body weight (formal question only)	16%	13%	391	1.8%	n.s.	-

[a]*Percentage of Q-A sequences that had to be excluded from all eligible Q-A sequences*
[b]*This question was only asked to respondents who ever walk 10 minutes on end*

For Q38 (perception of body weight), we manipulated both question wording and the types of alternatives. The percentage of mismatch answers for the four question versions is shown in Table 7-17. In accordance with hypothesis 2a, we found that conversational alternatives yield less mismatch answers than formal alternatives. However, the difference was only significant for the formal question wording. We did not find a significant effect in the percentage of mismatch answers for question wording, although the conversational versions, both for conversational and for formal alternatives, in accordance with hypothesis 1, generate more mismatch answers than formal questions.

Table 7-17 Percentage of mismatch answer for four versions of Q38

Conversational question	Perception of body weight (Q38)				
Q-A sequences	Formal alternatives *(too light, too heavy, not too light and not too heavy)*		Conversational alternatives *(too skinny, too fat or good)*		Total
With mismatch	33	**31%**	21	**21%**	
Without mismatch	73	69%	77	79%	
Total	106	100%	98	100%	204
χ^2= n.s. (3.8% Q-A sequences excluded)					
Formal question	Perception of body weight (Q38)				
Q-A sequences	Formal alternatives *(too light, too heavy, not too light and not too heavy)*		Conversational alternatives *(too skinny, too fat or good)*		Total
With mismatch	48	**25%**	27	**14%**	
Without mismatch	141	75%	165	86%	
Total	189	100%	192	100%	381
χ^2= 7.74** (4.3% Q-A sequences excluded)					

7.4.5 Effects of listed and implicit alternatives for choice questions (H2b, choice questions)

According to hypothesis 2b, implicit alternatives yield less mismatch answers than listed (either conversational or formal) alternatives. In order to test this hypothesis, five questions were manipulated. The results will be presented in two separate tables.

For questions Q33 (number of years since last visit to the General Practitioner) and Q34 (number of months since last visit to the Dentist's) we manipulated both question wording (conversational or formal) and wording of alternatives (implicit or formal). The percentages of mismatch answers for the four versions of each of these questions are shown in Table 7-18

Question Q33 (last visit to the General Practitioner) shows results contrary to hypothesis 2b: Implicit alternatives yield more mismatch answers (59% and 56% for conversational and formal questions respectively) than listed alternatives 26% and 39% for conversational and formal questions respectively).

For Q34 (last visit to the Dentist), in accordance with hypothesis 2b, implicit alternatives yield less mismatch answers than formal alternatives, but this difference is not significant.

The percentages of mismatch answers for Q33 and Q34 do not confirm hypothesis 1 (conversational versus formal question wording). The difference in the percentage of mismatch answers is only for Q33, within implicit alternatives, in accordance with hypothesis 1, in all other cases, the formal versions yield more mismatch answers than the conversational versions.

Table 7-18 Percentage of mismatch answer for four versions of Q33 and Q34

Conversational question	**Last GP's visit (Q33)**				
Q-A sequences	Formal alternatives *(shorter than a year ago, between one and two years ago or longer than two years ago?)*		Implicit alternatives *(How many years..)*		Total
With mismatch	44	**26%**	96	**59%**	
Without mismatch	127	74%	67	41%	
Total	171	100%	163	100%	334
χ^2 = n.s. (13.5% Q-A sequences excluded)					
Formal question	**Last GP's visit (Q33)**				
Q-A sequences					Total
With mismatch	41	**39%**	50	**56%**	
Without mismatch	65	61%	40	44%	
Total	106	100%	90	100%	196
χ^2 = n.s.(12.5% Q-A sequences excluded)					
Conversational question	**Last GP's visit (Q34)**				
Q-A sequences	Formal alternatives (shorter than a year ago, between one and two years ago or longer than two years ago?)		Implicit alternatives (How many years..)		Total
With mismatch	91	**50%**	78	**43%**	
Without mismatch	90	50%	103	57%	
Total	181	100%	181	100%	362
χ^2= n.s. (6.2% Q-A sequences excluded)					
Formal question	**Last GP's visit (Q34)**				
Q-A sequences					Total
With mismatch	60	**57%**	55	**52%**	
Without mismatch	45	43%	51	48%	
Total	105	100%	106	100%	211
χ^2= n.s. (5.8% Q-A sequences excluded)					

Questions Q16, Q21 and Q24 were manipulated with respect to the wording of alternatives. For Q16 (watching television) and Q21 (using breakfast) three types of alternatives were compared (conversational, formal and implicit alternatives), whereas for Q24 (non-alcoholic beverages) only formal and implicit alternatives were compared. Questions Q16 and Q21 are both behavioral frequency questions that ask for the number of days. In case of conversational and formal alternatives, four response categories are used, whereas for implicit alternatives eight categories are applied (i.e., a number between 0 and 7).

The results with respect to these questions are presented in Table 7-19. It appears that all three questions show different results. For Q16 (days watching television) results are opposite to hypothesis 2b. Although we found significant results for question Q21 (days using breakfast), the difference does not clearly support nor contradict the hypothesis that implicit alternatives yield less mismatch answers than listed alternatives . The result for Q21 is especially accounted for by a difference between formal and implicit alternatives, but not by a difference between conversational and implicit alternatives. For Q24, implicit alternatives also yield less mismatch answers than formal alternatives, but this difference is very small and not statistically significant.

Table 7-19 Percentage of mismatch answer for types of alternatives (listed versus implicit)

	Days watching Television (Q16)		
	(n = 372 Q-A sequences, 5.3% excluded)		
Conversational alternatives	*'every day', 'most days' 'some days; 'hardly ever'*		
Formal alternatives	*'practically 0 days', '1 to 3 days', 4 to 6 days', '7 days a week'*		
Implicit alternatives	*How many days?*		
	Conversational alternatives	Formal alternatives	Implicit alternatives
Conversational Q	29%	40%	**61%**
Conversational versus implicit χ^2= 24.66*, Formal versus implicit χ^2= 5.29*			
	Days using breakfast (Q21)		
	(n = 355 Q-A sequences, 14.0% excluded)		
Conversational alternatives	*Do you on weekdays never, once in a while, most days or every day use corn products such as bread, muesli or cornflakes as breakfast*		
Formal alternatives	*Do you on weekdays never, 1 to 2 days, 3 to 4 days or all 5 days use corn products such as bread, muesli or cornflakes as breakfast*		
Implicit alternatives	*How many weekdays do you use corn products such as bread, muesli or cornflakes as breakfast?*		
	Conversational alternatives	Formal alternatives	Implicit alternatives
Formal Q	44%	**62%**	52%
Conversational versus implicit χ^2= n.s., Formal versus implicit χ^2= 4.23*			
	Non-alcoholic beverages (Q24)		
	(n= 343 Q-A sequences, 16.9% excluded)		
Formal alternatives	*During a day, do you use more than 8 cups, about 8 cups or less than 8 cups of coffee, tea and other non-alcoholic beverages?*		
Implicit alternatives	*What is the total number of cups of coffee, tea and other non-alcoholic beverages that you usually use on a day?*		
	Formal alternatives	Implicit alternatives	
Formal Q	30%	28%	
χ^2= n.s.			

** p < 0.01, * p < 0.05

To summarize, hypothesis 2a could only be confirmed for choice questions that concerned behavioral frequency and perception towards body weight. Hypothesis 2b could not be confirmed. We therefore have to abandon this hypothesis.

7.4.6 Effects of questions requiring more or less cognitive processing (H3)

Hypothesis 3 concerned the effects of difficulty of question wording. We expected that questions requiring a large amount of cognitive processing (i.e., difficult questions) would yield more mismatch answers than questions requiring relatively little cognitive processing (i.e., easy questions).

To vary the difficulty of questions (see section 7.2.5), for the difficult version only one question was asked, whereas for the easy questions the same information was obtained by means of two questions in a row. These two questions in a row were always held constant for

other manipulations (i.e., a formal question was preceded by a formal one, and a conversational question by a conversational one).

In order to test whether problematic behaviors that are known to indicate question difficulty (other than mismatch answers) occurred, we analyzed the Q-A sequences of both question types for differences in the occurrence of such behaviors. It turns out that difficult questions yield more requests for clarification, more don't know answers, and more invalid answers than easy questions. This suggests that our manipulation of difficulty indeed had the intended effect. However, the number of occurrences of our indicators was too small to yield significant results.In Table 7-20 the percentage of mismatch answers in Q-A sequences concerning difficult questions and easy questions is shown. As the results in Table 7-20 show, our hypothesis cannot be confirmed. There is no significant difference between the difficult and easy question wording in the percentage of mismatch answers for either of the questions.

Table 7-20 Difference in percentage of mismatch answers for difficult and easy choice questions

	Question wording		N	% missing[1]	$\chi 2$
Question	Difficult	Easy			
Q12 Hours/Minutes sports					n.s.
Non-ambiguous, formal Q	28%	19%	100	8.1%	n.s.
Q14 Days occupied with walking					n.s.
Conversational Q	47%	41%	255	7.5%	n.s.
Formal Q	34%	37%	248	6.6%	n.s.

[1] *Percentage of Q-A sequences that had to be excluded from all eligible Q-A sequences*

As we already indicated in section 7.4.3, the number of respondents that replied to the sports question (Q12) is rather low due to a preceding filter question. Furthermore, the question was manipulated for three different hypotheses, thus we could only compare versions that were held constant for the other manipulations. However, Q14 (Days occupied with walking) did not suffer from these biases, as about 95% of the respondents did answer this question, but did not confirm hypothesis 3 either.

Hypothesis 3 thus could not be confirmed. Our manipulation, that comprised asking the same information in a single question or in two questions did not create any differences in the occurrence of mismatch answers. However, the occurrence of other problematic respondent behaviors indicates that respondents may indeed have experienced the difficult questions as more difficult than the easy questions.

7.4.7 Effects of ambiguous and non-ambiguous questions (H4)

Hypothesis 4 concerned the difference in the occurrence of mismatch answers for questions containing ambiguous concepts (i.e., ambiguous questions) and questions not containing ambiguous concepts (i.e., non-ambiguous questions).

Table 7-21 shows the percentage of Q-A sequences with mismatch answers for ambiguous and non-ambiguous questions. Hypothesis 4 can be confirmed for only one question (Q1; perception of health). The ambiguous question yields mismatch answers in 14% of the Q-A sequences, whereas the non-ambiguous version yields mismatch answers in only 8% of the Q-A sequences.

Table 7-21 Difference in percentage of mismatch answers for ambiguous and non-ambiguous choice questions

	Question wording		N	% missing[1]	$\chi 2$	exp
Question	Non-ambiguous	Ambiguous				
Q1 Perception of health	8%	14%	575	5.7%	3.94*	*
Formal Q	8%	13%	288		n.s.	+
Conversational Q	8%	15%	257		n.s	+
Q10 Engaged in sports						
Formal Q	12%	4%	289	7.6%	n.s.	-
Conversational Q	2%	3%	276	7.2%	n.s.	+
Formal Q (corrected)	1%	4%	270	14.2%	n.s.	+
Conversational Q (corrected)	2%	3%	276	7.2%	n.s.	+
Q12 Hours/minutes sports						
(Formal, difficult version)	28%	9%	318[a]	8.6%	n.s.	-
Q17 Time watching television						
Formal Q	28%	22%	195[b]	8.7%	n.s.	-
Conversational Q	22%	28%	337	9.3%	n.s.	+
Q40 Level of education						
Formal Q	22%	21%	279	8.5%	n.s.	+
Conversational Q	20%	20%	297	2.6%	n.s.	0
Q43 Ownership of car						
Formal Q	28%	3%	484	3.0%	n.s.	-

[a] *This question was only asked to respondents who had been engaged in sports the last 12 months*
[b] *This question was only asked to respondents who watch television at least once a week*
[1] *Percentage of Q-A sequences that had to be excluded from all eligible Q-A sequences*
*** p < 0.01, * p <0.05*

The strongest difference contradicting our hypothesis could be found for question 43. The non-ambiguous wording (‘Including al licensed vehicles, do you or anyone in your household own a car or motorcycle?’) appeared to create more ambiguities than both the ambiguous conversational wording (‘Do you, or anyone at your home have a car?’) and the ambiguous formal wording (‘Do you, or anyone in your household own a car?’). Moreover, in this case the manipulation did not control for the number of words, thus ambiguity and question length were confounded variables. The ‘non-ambiguous’ question not only yielded more mismatch answers than the ambiguous versions, but also more requests for clarifications (occurring in

6% of the Q-A sequences for the formal non-ambiguous version and 1% of the Q-A sequences for both the formal and the conversational ambiguous versions, χ^2=14.26, p <0.01).

Also question 10 (engaged in sports) and question 12 (hours and minutes spent on sports) yield results contrary to our hypothesis. It turns out that differences found *within* the formally worded questions contradict the hypothesis that non-ambiguous questions yield less mismatch answers than ambiguous questions. However, for question 10 this appears to have been caused by the question order effect we also described in section 7.4.3. If we delete Q-A sequences with mismatch answers due to the question order effect, there is no significant difference in the percentage of mismatch answers for the ambiguous and non-ambiguous question. For Q17 and Q40 no significant effects were found.

To conclude, hypothesis 4 could only be confirmed for one question, which seems to support Mallison's (2002) finding that respondent's may use different frames of reference in answering the original wording 'Do you have a very good, a good, reasonable or bad health?'. Our 'non-ambiguous' rewording, i.e., asking respondents to take their own age as a frame of reference yielded less mismatch answers than the original 'ambiguous' wording. Nevertheless, for the other questions we could not find support for our hypotheses. In case of the question concerning car ownership, we even found a very large difference in the opposite direction. In this case the non-ambiguous version of the question contained so much specifications that it increased the complexity of the question. Adding specifications may be helpful in case a question can clearly be interpreted in different ways, depending on the respondent's frame of reference, but may also increase the complexity of the question, thus creating rather than solving ambiguity.

7.5 Discussion

The goal of this experiment was to confirm hypotheses about the occurrence of mismatch answers. One reason for the occurrence of mismatch answers is that people are used to participate in ordinary conversations, and apply their style of responding to survey interviews. When survey questions resemble expressions commonly used in ordinary conversations, respondents will not be focused on the task of giving precisely formatted answers, yielding a high number of mismatch answers. A formal question on the other hand, triggers respondents to focus adequately on the task of formulating precise answers. In our experiment, in some cases we could find support for hypothesis 1, but in general the comparisons of conversational and formal questions did not yield a clear difference in the percentage of mismatch answers. With hypothesis 2a we aimed to test the effects of the conversational character of response alternatives. Respondents are not accustomed to use formal words, and as a consequence give mismatch answers. When response alternatives are used that are formulated according to language in ordinary conversations, respondents will have less difficulties with using such conversational alternatives, and as a consequence give less mismatch answers than when formal alternatives are used. We could confirm this

hypotheses 2a for the majority of the questions tested. For a few questions we could not confirm this hypothesis, but for some questions we found strong effects.

Hypothesis 2b, implicit alternatives yield less mismatch answers than listed (either conversational or formal) alternatives, could not be confirmed.

Next to habits of ordinary conversations, a second reason for the occurrence of mismatch answers is that information required to arrive at an answer to a survey question is often not readily available. Retrieval of information is likely to be verbally expressed, which increases the chance of mismatch answers. Only two questions were manipulated to test this hypothesis, which could not be confirmed. It is possible that our manipulations did not have the effect that we expected.

A third reason for the occurrence of mismatch answers is task uncertainty: respondents have difficulties in translating detailed information into the response categories. Only one of the seven questions showed support for this hypothesis. The results for other questions showed that decreasing the ambiguity can increase the chance of mismatch answers due to the increased complexity of questions.

Unintended effects of the manipulations may explain the fact that the hypotheses 1, 3 and 4 could not be confirmed for the majority of the questions, but it is also possible that alternative explanations or possible interaction effects were present. These will be discussed in the next sections.

7.5.1 Success of the manipulations

A possible reason for our lack of finding effects of question wording might be our operationalization of the manipulations. For example, it might have been necessary to use more extreme manipulations to be able to confirm our hypotheses.

We were primarily concerned with the use of feasible and realistic survey questions. Therefore we used assertions that were derived from actual surveys. Our priority given to external validity may have sacrificed internal validity of the study, as the differences between the questions may not have been extreme enough. In spite of this, internal validity concerns were also a reason to use realistic survey questions. Extreme manipulations might have alerted interviewers of the experimental character of the study. In this study it was very important that interviewers were unaware of the actual hypotheses being tested. Their knowledge of the expected outcomes could have influenced their behavior in the interaction with the respondent. For example, they could have stressed the importance that respondents formulate their answers as precisely as possible, and they could have done this in different ways for different versions of the same question. In that way it would have been impossible to distinguish effects of interviewer behavior from effects of question wording. Fortunately the interviewers' behavior did not indicate any suspicion with regard to the goals of the study. We could not find any differences in interviewer behavior, related to the different versions of the questions. Moreover, interviewers were surprised when they were, at the end of the study, informed about the experimental character of the study.

Manipulation of conversational character of questions and alternatives

The manipulations of the conversational character of questions and alternatives were based upon word frequencies in ordinary conversations. This strategy was most feasible for the wording of alternatives. For the manipulation of questions it may have been more problematic to use word frequencies. Many survey questions are not likely to be asked in such a way in ordinary conversations, even if they mainly consist of common words. Probably, questions in ordinary conversations are differently structured as compared to formal questions. The grammatical structure of questions was kept equal across conditions. Nevertheless, it is very well possible that the grammatical structure of both our formal and conversational questions, are quite formal, compared to the language as used in common conversations. A better strategy would have been to base the manipulations of question wording on actual frequencies of grammatical structures or complete sentences, instead of word frequencies. However, to our knowledge a frequency database with such information does not exist.

Manipulation of difficulty

For the comparisons of 'easy' and 'difficult' questions, only two questions were manipulated. Although there are indications that the 'difficult' questions indeed were experienced as more difficult than the 'easy' questions, the questions did not differ in the number of mismatch answers they yielded.

It turned out to be not so straightforward to create questions that collect the same information but differ with respect to difficulty. The difficulty of a question is much more related to the character of the information that is asked than to the wording and structure of the question. A decomposition strategy, asking several questions requiring cognitive processing in small steps, rather than one question, did not decrease the number of mismatch answers.

For the sports question we suffered from the problem that respondents who usually produce most mismatch answers (i.e., the older ones) did not answer this question when they indicated that they had not been engaged in sports. Furthermore, one of the questions was manipulated in order to test three different hypotheses. However, for the question that did not suffer from these biases we could not confirm the hypothesis either. Thus, it is likely that the manipulation of difficulty was not successful. The decomposition of questions did not help respondents to separate the retrieval of information required to answer the question.

Manipulation of ambiguity

In general, the questions with ambiguous concepts did not yield more mismatch answers than the questions with less ambiguous concepts. The manipulations were focused on the specification of ambiguous concepts: in ambiguous questions concepts such as 'car', 'health', 'watching television' and 'completed education' were not specified, whereas in non-ambiguous questions they were specified (e.g., explaining what comprises a car and that the

judgment of own health must be compared to health of peers). Creating non-ambiguous questions had the effect that some questions contained so many specifications that they evoke more confusion than the ambiguous equivalents. So, the advice of the 'question appraisal system' (Willis and Lessler 1999) to avoid undefined common terms was not useful with respect to the effect on the occurrence of mismatch answers.

Another problem with finding effects of ambiguity is also that respondents may not always reveal their problems. So even when a question that is formulated as ambiguous, is indeed perceived as ambiguous, then still only a small part of respondents may give a task mismatch answer.

7.5.2 Compatibility of question wordings

To test whether questions that differ with respect to their conversational character, difficulty or ambiguity affect the occurrence of mismatch answers, they should have equivalent meanings. When questions differ with respect to their meaning, it is difficult to distinguish effects from this difference in meaning from effects of manipulations. In that case, replacement of the question that yielded most mismatch answers with the question with least mismatch answers will also yield measurement of different concepts.

A comparison of response distributions may be used as an indicator of which versions of the same question may have differed in meaning. However, such a comparison is problematic for two reasons. Firstly, when no difference in response distributions is found, this is no guarantee that questions have the same meaning. Secondly, a difference in response distributions may also be the result of the occurrence of mismatch answers or other problematic deviations. In an earlier study (Dijkstra and Ongena forthcoming) it was shown that the occurrence of problematic respondent behaviors yielded a lower response validity. Therefore we can only compare the response distributions for Q-A sequences during which no problematic deviations occur. However, this comparison is also complicated: when we include only Q-A sequences during which no problematic deviations occur, we may have a sample of respondents that differs from the sample of respondents that produced problematic deviations. The occurrence of problematic deviations is likely to differ among respondents and may be related to respondent characteristics.

Nevertheless, we compared response distributions for the different question versions. This comparison showed that for most questions there were no differences in the response distributions between different versions. For the few questions that did show a difference in response distribution, we tried to explain this in terms of a difference in the meaning of the question versions. However, we were not able to find clear indications of such differences.

7.5.3 Consequences for survey practice

The results of this study showed that conversational alternatives decrease the chance of mismatch answers. Therefore, especially in case of assertions, conversational alternatives should be used. Our results did not clearly show effects of question wording. In some cases,

formal question wording, appeared to decrease the chance of mismatch answers, but there is no clear evidence of this effect. The effects of alternatives and questions were not systematically compared for all questions. We did not formulate a hypothesis on the effects of interaction between question wording and the types of alternatives. In a multilevel logistic regression on the Q-A sequences of all assertions no overall interaction effects were found between effects of question wording and types of alternatives. However, in a few cases, a different effect of question wording showed up, depending on the types of alternatives used.

When a conversational question is accompanied by formal alternatives, respondents will not use the formal wording of the alternatives and as a result give mismatch answers. When a conversational question is accompanied by conversational alternatives, the conversational answers of respondents are more likely to match the conversational alternatives (i.e., respondents give less mismatch answers). A formal question is assumed to trigger adequately formatted answers in all cases, yielding less mismatch answers than conversational questions, both in case of conversational and formal alternatives. Thus, we would expect strongest effects of question wording within formal alternatives.

For the 'Public Health' assertions an effect of question wording was found within conversational alternatives: conversational assertions yielded more mismatch answers than formal assertions, but this effect was not found within the formal alternatives So, this is contrary to our expected interaction effect. However, we can assume that not only wording of the question proper, but also the response alternatives trigger a style of responding, which may have resulted in an opposite interaction effect. The conversational style of responding that is triggered by conversational questions *and* alternatives not only triggers respondents to use conversational words, but may also cause them to perform other conversational behavior, i.e., being less precise, and to elaborate. When respondents are confronted with conversational questions and response alternatives, they have not received any signal about the formal character of the survey, and may start to elaborate and forget to give any adequate answers at all. When respondents are provided with a formal question with conversational alternatives, they received signals about the formal character of the survey and at the same time have response options that they are accustomed to use.

Our findings suggest that, to reduce the chance that mismatch answers occur, conversational response alternatives should be used. When, in addition, respondents are triggered to respond in a formal rather than a conversational style, they are most likely to give precisely formatted answers without distracting elaborations.

Based upon the assumption that mismatch answers (and other problematic deviations) decrease the response validity, we assume that the conversational alternatives yielded highest response accuracy. Unfortunately, we do not have validity measures available to test this assumption. We can only compare Q-A sequences with and without mismatch answers for differences in response distributions and differences in correlations between variables. However, it is also possible that the occurrence of mismatch answers is related to the intended answers: respondents for whom a particular response is applicable, may also be more likely to produce a mismatch answer. Furthermore, the effect of the occurrence of

mismatch answers on response validity highly depends on the interviewer's behavior. For example, interviewers may infer intended answers from mismatch answers, and use this inference to suggest an alternative or directly score the inferred response without verification. A clear-cut consequence of the occurrence of mismatch answers is the fact that it extends the interaction. The number of events in a Q-A sequence with a mismatch answer is on average significantly higher (M=7.6 events) than the number of events in a Q-A sequence without a mismatch answer (M = 4.4 events, $t = -52.49$, $df = 24440$, $p < 0.01$).

8 Summary and discussion

8.1 Main findings

The goal of this thesis was to gain more insight in the interactional process of questioning and answering in the survey interview. This process takes place in so-called question answer sequences (Q-A sequences). A Q-A sequence consists of all utterances that concern a question, starting with the interviewer asking a question and ending just before the next question from the questionnaire is posed.

A 'paradigmatic' Q-A sequence (Schaeffer and Maynard, 1996) is perfect from a survey researcher's point of view. In such a sequence the interviewer poses the question as scripted and the respondent gives an appropriate answer immediately. When the exchange of information does not proceed as a 'paradigmatic sequence', the course of the Q-A sequence will affect the eventual answer and response errors may occur. For example, respondents often give answers that are not exactly formatted according to the response alternatives (mismatch answers). As a consequence, interviewers may suggest a particular alternative to the respondent. In that case, the eventual answer of the respondent is quite probably affected by this suggestion. When such influences occur, the quality of the data collected may be affected negatively.

Of course the presence of paradigmatic Q-A sequences is no guarantee for the absence of response errors. For example, respondents may give socially desirable answers or misunderstand questions, without causing deviations from the paradigmatic sequence, and thus nothing can be detected by means of the analysis of the verbal behavior in Q-A sequences. However, deviations from the Q-A sequences provide overt indicators of problematic processes in answering questions. An inspection of the types of problems that occur in these verbal deviations can offer insight into causes and consequences of such problems and the consequences for data quality.

In this thesis we aimed to answer four questions, which will be discussed in the next subsequent sections.

8.1.1 Problematic behavior of interviewer and respondent: theoretical viewpoints

The first research question, concerning which types of problems in the interaction between interviewer and respondent can be expected from different theoretical explanations was dealt with in chapter 2. In this chapter we provided an overview of conversational and cognitive theories that are relevant to the interaction in the survey interview. The cognitive processing of answering a survey question, which can be described by means of four steps (Tourangeau et al. 2000), may influence the course of the interaction through conversational principles.

For example, performing step 1, question comprehension, can create problems for respondents when the question's meaning is not clear to them. If and how problems in understanding the question's meaning are expressed by respondents and how interviewers subsequently handle this will depend on conversational principles. The more respondents avoid face threats (i.e., are polite and avoid to insult the other person), or pursue a satisficing

strategy (i.e., invest less effort in their task), have low task involvement (i.e., the motivation to approach the interview seriously), and have experience with 'What ever it means to you' replies (i.e., standardized 'clarifications' of questions), the less likely respondents will explicitly request clarification of the meaning of the question. In the same way, the more interviewers avoid face threats, commit themselves to standardization rules, or are unable to recognize the source of the problem in understanding, the more they will avoid explicit clarification of the question. In case interviewers do not follow standardization rules they may probe suggestively, or infer respondents' answers.

Steps 2 and 3, retrieving relevant information and forming a judgment from this information may be visible in the interaction in case respondents perform an enumeration strategy. This enumeration may be verbally expressed before an answer is given. Interviewers may respond to such enumerations before the respondent has come up with a final answer. This response may comprise an inference, which is of course problematic for the quality of the response obtained. Furthermore, respondents may provide cues, such as hesitations, verbally express uncertainty about the adequacy of their answer, or give an imprecise response (a mismatch answer). Again, the more interviewers avoid face threats, the more they will avoid probing for precise answers, and just infer respondents' answers.

Step 4, formatting the response, may affect the interaction according to a conversational principle, called the preference for agreement. Respondents are likely to format a response initially in a preferred agreeing format, or start with hesitations before they give their disagreeing answer. When interviewers accept any response too quickly, it is possible that respondents never restate such initial responses into the dispreferred format. Furthermore, respondents may view the survey as an ordinary conversation. This may cause them to elaborate their answers, and give conversational answers. A conversational answer is likely not formatted according to the required response format. In short, the theoretical models presented in chapter 2 show that the course of the interaction can take on a variety of forms. Both conversational and cognitive aspects may be responsible for quite a lot of different overt problems, like requests for clarification, mismatch answers, improper question reading, suggestive behavior, and so on. Such problems will quite likely affect the course of the interaction and in turn may yield new problematic behavior.

8.1.2 Methods to identify interactional problems in survey interviews

The question how courses of interactions can be analyzed in a systematic way, to test relations between behaviors, was addressed in chapter 3. In order to analyze Q-A sequences in a quantitative way, it is necessary to code the verbal behavior of interviewers and respondents. The basis of the procedure is to systematically assign codes to behaviors in a Q-A sequence. The characterization included in the codes can be a pure description of the kind of behavior, such as 'interviewer poses question', but it can also include an evaluative component, for example evaluating the adequacy of an instance of behavior, according to standardization rules (e.g., 'interviewer poses question as worded' and 'interviewer poses question inadequately'). Usually multiple coders are involved in the behavior coding

procedure, and therefore the inter-coder (but also intra-coder) reliability are important issues.

In chapter 3 we also gave an overview of different methods of behavior coding. We were found 48 different coding schemes. This overview showed different procedures and strategies that have to be decided on. A first decision is the selective character of the scheme. All utterances can be coded (full coding), or only a selection of important behaviors in view of specific research questions (selective coding). A second decision, which partly depends on the first, is the unit of analysis. Coding can take place at the level of the utterance, the exchange level or the whole Q-A sequence. Next decisions concern preservation of sequential information, practical procedures (coding live, from tapes or with transcripts), and the type of coders used (interviewers, the researcher, or specially trained coders).Which coding strategy is used will have consequences for the types of analysis that can be performed. Schemes that aim for quick results may retrieve sufficient information from frequency analyses. This may be the case for behavior coding studies in a pretesting phase or during data collection, when relevant parts of data collection are monitored and quick feedback is required. Such schemes are limited to selective schemes (with less than 15 codes). In case of an evaluation of the data collection process, quick results may be less important. However, a detailed explanation of causes of problematic behaviors is usually not the kind of information that is sought for in these studies, and therefore selective coding schemes with a slightly higher number of codes (i.e., around 20) may be appropriate.

In case of exploratory analyses of the interaction, detailed information is required, and full coding schemes with sequential information seem most appropriate. For a practical application of such schemes, software like the Sequence Viewer program is available (see Dijkstra 2002).

Because our research questions are focused at a complete description of the interaction, and also refer to the order of occurrence of behaviors, we thus need a full coding scheme with preservation of sequential information. This is by far the most labor-intensive, but also the most informative method, as a lot of information can be derived from sequence analyses.

The coding scheme that we used for this method is described in chapter 4. This scheme met our criteria of feasibility and the required amount of detail included in the codes. With this multivariate coding scheme (Dijkstra 1999), behaviors in the interaction are coded on a number of different coding variables. Each variable describes a particular aspect of the utterance. The combination of values yields a code string that constitutes a meaningful description of the utterance. The multivariate character of the scheme also makes it more flexible to switch from rough analyses (using only part of the coding variables) to detailed analyses (using most or all variables). Furthermore, the scheme is relatively easy to use by coders. They do not have to choose from only a long list of codes, but instead choose from a few categories for each coding variable.

A comparison of codes across the 48 different coding schemes showed that with our coding scheme not only nearly all categories that are covered by other schemes can be coded, but also many more. It was illustrated how the coding scheme can be used to describe almost any kind of verbal behavior that is relevant to the course of the interaction. In summary, coding procedures and strategies, as well as the actual codes themselves, highly depend on

the goal or focus of the study. As our study is focused at the description of the course of the interaction between interviewer and respondent, we decided to use a very detailed, full coding scheme, using the utterance as unit of analysis and preserving sequential information.

8.1.3 Causes of problematic deviations from the paradigmatic Q-A sequence

The third research question, concerning the most frequent occurring problematic deviations and their causes, was answered in chapter 5. In order to answer this question, we used telephone interviews of a survey about behaviors and opinions concerning watching television and commercials. The exploratory analyses of the transcribed and coded interviews showed that in almost 50% of the Q-A sequences problematic deviations occurred.

The respondent was usually the first to produce problematic deviations in a Q-A sequence. Mismatch answers occurred most frequently, especially as the first problematic deviation in a Q-A sequence. In case of a mismatch answer interviewers are required to probe for an adequate answer, which frequently causes interviewers to probe suggestively or perform another type of inadequate behavior. Apparently, it is most important to train interviewers how to deal with mismatch answers.

An even better way to improve data quality is trying to prevent the occurrence of mismatch answers. Question wording and the type of alternatives used play an important role here. In general, response alternatives that do not correctly match the question generate a high number of mismatch answers. Response alternatives with a four or five-point Likert-type scale also yield a high number of mismatch answers. Furthermore, we found that imprecise interviewer instructions on how to repeat alternatives for each question in a battery of questions with the same response alternatives, increases the chance of mismatch answers. In contrast, yes-no questions yield a low number of mismatch answers, at least, if the response alternatives do not include other alternatives than ‘yes’ and ‘no’ (or, if applicable ‘don’t know or ‘refuses to answer’).

Causes of other problematic deviations were also found to be related to question characteristics. For example, general questions after specific questions cause interviewers to skip the general question, or to probe insufficiently. Furthermore, older, lower educated and female respondent produced more problematic deviations than younger, higher educated and male respondents.

To summarize our answer to the third research question, mismatch answers are by far the most frequently occurring problematic deviations, and moreover an important cause of problematic behavior of the interviewer. Question wording, but especially the kind of response alternatives, in addition to respondent characteristics seem to be the most important causes of mismatch answers.

8.1.4 Theoretical explanations for the occurrence of mismatch answers

The frequent occurrence of mismatch answers, and the fact that they are also the most important cause of problematic interviewer behavior, motivated us to further study possible

causes of this problematic deviation, in order to answer our fourth research question. In chapter 6, three reasons for the occurrence of mismatch answers were discussed.

Firstly, a conversational problem may occur. We assume that this is the most important cause of mismatch answers. Respondents usually do not have a clear idea about what is expected of them. They confuse the standardized survey interview with an ordinary conversation, and thus format their answer to the survey question in the same way as they are accustomed to answer questions in ordinary conversations.

For example, when respondents are asked how many days a week they watch television, they may think that it is acceptable to give an answer like "Most days" instead of exactly defining the number of days. However, such an answer is not directly codable by the interviewer because it does not match one of the fixed alternatives. These mismatch answers are called conversational mismatch answers.

Secondly, ambiguity of question meaning may cause task uncertainty. Although respondents may have all relevant information available, they face the problem to translate their specific situation into one of the response alternatives offered. As a result, respondents will give considerations, and increase the chance of providing a mismatch answer. These mismatch answers are called task mismatch answers.

Thirdly, a cognitive problem may occur, when the information required to answer a question is *not* readily available in the respondent's memory. Respondents may in that case start to think aloud, and also give verbal considerations before their answer, increasing the chance of a mismatch answer. These mismatch answers are called cognitive mismatch answers.

We formulated a number of hypotheses about effects of particular wordings of questions and response alternatives on the probability of occurrence of mismatch answers. These differences in wordings were based upon the hypothesized causes of the three types of mismatch answers discussed above. In the next sections we will discuss these relations in more detail, and summarize the results of a non-experimental survey and an experimental survey.

The non-experimental survey (chapter 6) concerned interviews from the Dutch pilot study of the European Social Survey (ESS). The questionnaire of this CAPI survey comprised 268 different questions, of which several question categories could be distinguished in order to test our hypothesis non-experimentally. In the experimental survey (chapter 7), we used question wordings that were derived from actual surveys concerning health issues, and created multiple versions of the same question, to compare the effects of question wording and types of alternatives.

8.1.5 Conversational mismatch answers: Question wording

We hypothesized that 'conversational' questions (i.e., formulated in a manner that is common in ordinary conversations) will misleadingly signal respondents that a conversational style of responding, usually less exact and precise, is appropriate. In contrast, formal questions signal

respondents that an exact and precise response is required. Hence, we hypothesize that conversational questions generate more mismatch answers than formal questions.

We were able to confirm this hypothesis in the non-experimental analysis of the ESS data for questions without show cards. Questions from this survey that were rated as 'conversational' yielded more mismatch answers than questions that were rated as 'formal'. However, in the experimental study, the hypothesis could only be confirmed for some opinion assertions, and some background questions, but in general effects of the conversational character of questions were not found. For a few questions the results even appeared to contradict the hypothesis.

In the non-experimental study, the effect of the conversational character of questions was found mainly for difficult questions (for example behavioral frequency questions or retrospective questions), but for easy questions the effect was not necessarily present. In the experimental survey such a difference was not found.

We assume that our manipulations of the conversational character of questions in the experimental survey were not extreme enough. The operationalization of conversational questions was based upon composing questions with common words (i.e., with high frequencies in a word frequency database of Dutch conversations). With our manipulation, we also aimed to use realistic survey questions. Hence we had neither created extremely conversational question versions, nor extremely formal versions. In fact, many of the conversational questions used can still be considered as unlikely to occur in ordinary conversations. Thus, in some cases, we were in a sense comparing 'formal' questions with 'more formal' questions. Manipulations by means of frequencies of complete sentences (i.e., taking also a conversational grammatical structure, and likelihood of being used in conversations into account) might have yielded larger effects.

8.1.6 Conversational mismatch answers: type of alternatives

Instead of formulating questions in a formal way to prevent conversational mismatch answers, we can also use conversational response alternatives in order to prevent that respondents have to answer questions in a way they are not accustomed to. Thus, alternatives that are frequently used may decrease the chance of mismatch answers. Formal words on the contrary, are less often used in ordinary conversations. Hence, we hypothesized that questions with conversational alternatives yield less mismatch answers than questions with formal alternatives.

This hypothesis could be confirmed both in the non-experimental and the experimental study. In this case, using an operationalization of response alternatives based upon word frequencies in ordinary conversations was an effective strategy. In the experimental study, strongest effects were found for assertions. Respondents typically treat assertions as yes-no questions (i.e., answering them with 'yes' or 'no'), and 'yes' and 'no' are also typically conversational responses (i.e., with a high frequency in the word frequency database).

For the assertions, we compared two- and three-point scales of yes-no alternatives (i.e., conversational alternatives) with a Likert-type scale, and with two- and three-point scales of

formal response alternatives. In all comparisons, yes-no alternatives yielded less mismatch answers than the formal versions. Least mismatch answers occurred when, with the conversational alternatives, a middle alternative was also used (i.e., 'maybe'). However, this effect could be related to the specific character of assertions (i.e., health perception assertions) and needs to be tested for opinion assertions, and for other topics as well.

There is a point of concern about the validity of 'yes' and 'no' as conversational response alternatives. As Houtkoop-Steenstra (2000) notes, 'ja' (i.e., 'yes') in Dutch is highly ambiguous. As a response alternative, 'ja' is intended to be used as an agreeing response. However, 'ja' may also be used as an acknowledgement token (like 'yeah' in English), and it may even be used as the beginning of a non-agreeing action (like 'well' in English). The meaning of 'ja' at the beginning of a turn becomes clear when the speaker continues. Therefore, as Houtkoop-Steenstra puts it, it is important that interviewers are careful enough "not to immediately treat a respondent's turn-initial 'ja' as an agreement or a confirmation" (p. 124).

From these concerns, one may wonder whether the low occurrence of mismatch answers, which we found for conversational alternatives that include 'yes', could be explained by those ambiguous instances of 'yes'. However, coders were explicitly instructed to code 'ja'-answers only as mismatch answers when they appeared to be intended as answers by respondents. Furthermore, the answer distributions did not indicate that the formal equivalents (such as 'true' and 'agree') were chosen less frequently than the conversational alternative 'ja'.

Questions can be accompanied by explicitly listed alternatives or by implicit alternatives. Questions of the latter type have an open-ended response format that implies a range of alternatives, e.g., a number of hours or minutes, a percentage, number of days, etc. However, such alternatives are not explicitly mentioned by the interviewer. In ordinary conversations, it is very unusual to mention response alternatives. Thus, we assumed that implicit alternatives are more conversational than listed alternatives, and as a consequence, questions with implicit alternatives are hypothesized to yield less mismatch answers than questions with listed alternatives.

This hypothesis could neither be confirmed by the ESS data, nor with the experimental data. Question versions with implicit alternatives did not yield less mismatch answers than question versions with listed alternatives.

8.1.7 Task mismatch answers

When questions contain ambiguous concepts this may cause task uncertainty. As a result of this uncertainty respondents may have trouble to decide between the alternatives, and are more likely to give a mismatch answer. We hypothesized that questions with ambiguous concepts that are not specified, yield more mismatch answers than questions in which concepts are specified. We were able to confirm this hypothesis with the ESS data. Questions

from this survey that were rated as 'ambiguous' yielded more mismatch answers than questions that were rated as 'non-ambiguous'. However, the difference in the percentage of mismatch answers was very small. Moreover, we could not find evidence in the interactions that the mismatch answers that occurred with ambiguous questions were due to task uncertainty. In the experimental survey, we could confirm this hypothesis for only one of seven questions. One question clearly showed results contrary to our expectations. The non-ambiguous version of this question contained so many specifications (about what can be considered as a 'car') that more uncertainty is created than with the question in which concepts are not specified. Thus, our manipulation of ambiguity had other effects than we intended. Furthermore, task mismatch answers do not occur as frequently as conversational mismatch answers, and thus effects of a decreased chance of task mismatch answers are difficult to find.

8.1.8 Cognitive mismatch answers

Questions that require substantive processing to arrive at an answer may cause state uncertainty. Respondents may have difficulty in retrieving information necessary to answer the question. This difficulty may be expressed by means of verbal considerations, and these considerations are likely to result in mismatch answers. Thus, we hypothesized that questions requiring information not readily available in memory (i.e., difficult questions) will generate more mismatch answers than questions requiring relatively little cognitive processing (i.e., easy questions). This hypothesis could be confirmed with the ESS data. Questions from this survey that were rated as 'difficult' yielded more mismatch answers than questions that were rated as 'easy'. Especially mismatch answers that were categorized as cognitive mismatch answers accounted for this effect.

However, in the experimental study, the hypothesis could not be confirmed. For comparisons across 'easy' and 'difficult' question versions, only two questions were manipulated. Although respondents' requests for clarification and 'don't know' answers indicated that the 'difficult' versions indeed were experienced as more difficult than the 'easy' versions, the questions did not differ in the number of mismatch answers they yielded.

It turned out to be not so straightforward to create question versions that collect the same information but differ with respect to difficulty. A decomposition strategy, asking several questions requiring cognitive processing in small steps, rather than one question, did not decrease the number of mismatch answers. Our finding seems to support those of other studies, in which similar decomposition strategies are judged as unlikely to increase accuracy of behavioral frequency reports when frequent and common behaviors are involved (Belli et al. 2000). Furthermore, it is possible that the questions were not decomposed in a way that was helpful for respondents. As the manipulated questions generally comprised questions on retrospective data, other techniques to create easier questions, especially those focused on aided recall, may be more effective to reduce the chance of mismatch answers (e.g., see Schwarz and Oyserman 2001).

Our answer to research question 4, which theoretical explanations can be found for the occurrence of problematic deviations, i.e., mismatch answers, is mixed. We hypothesized that respondents tend to answer questions in a survey interview in a way they are accustomed to in common conversations, yielding mismatch answers if these responses do not fit the prescribed alternatives. We further hypothesized that conversationally worded questions trigger a conversational way of responding. This hypothesis was confirmed in a non-experimental survey, but did not receive much support in an experimental survey, manipulating question wording. We also hypothesized that alternatives that consist of conversational words would yield less mismatch answers than formal alternatives. This hypothesis was generally supported both by the non-experimental and experimental study. Two other causes of mismatch answers, related to state and task uncertainty were hypothesized to be affected by the difficulty and the ambiguity of the question. Although support for both hypotheses was found in the non-experimental study, they could not be confirmed in the experimental survey.

8.2 Suggestions for further research

Although the results of this thesis provided answers to our four research questions, our answers are not definitive in all cases, and evoke a new series of questions.

8.2.1 Replication of results in other languages and other situations

The empirical results described in this thesis all concern surveys that were held in the Dutch language. It may be difficult to generalize the results to other languages, especially those concerning the conversational character of questions and alternatives. However, other studies have shown that the problems that arise as a result of the difference between ordinary conversations and standardized interviews are a universal phenomenon. Nevertheless, languages differ in their flexibility of creating sentence structures that fulfill specific requirements of survey questions, and are not awkward at the same time. For example, survey questions require that the question delivery component (which makes it possible to infer the meaning of the question) be placed at the end of the question, in order to avoid interruptions. Fulfilling this requirement is more easily done in Dutch than it is in English.

Furthermore, the empirical results of the experiment concern telephone surveys. It remains to be seen how these results can be generalized to face-to-face settings. In a face-to-face interview, non-verbal cues can play an important role in the interaction. Respondents may non-verbally signal difficulties they have with questions, and they may even answer by means of a nod.

Another issue is that in face-to-face interviews show cards can be used to decrease the chance of mismatch answers. Other studies have shown that questions with show cards yield less mismatch answers than questions without show cards (Dijkstra and Ongena forthcoming; Prüfer and Rexroth 1985; Sykes and Collins 1992). Show cards may, like formal question wording, trigger respondents to focus adequately on the task of formulating precise answers. Thus, the beneficial effect of show cards may be largely due to a decrease in the number of

conversational mismatch answers. However, with the data of the ESS pilot study we were unable to confirm this effect. In the ESS, show cards were used particularly for assertions, which yield a high number of mismatch answers in case the traditional Likert-type scale is used.

8.2.2 More response alternatives

In the experiment, the conversational alternatives used for the assertions comprised lists of no more than three alternatives. It would be worthwhile to study the effects of longer lists of conversational alternatives, because a researcher may wish to have measurements with more than three categories. This may especially be the case for attitude questions. With only two or three categories, a researcher is not able to measure attitude strength, restricting the measurement to attitude direction. Krosnick and Abelson (1992) argue that whenever attitude direction is measured, attitude strength should be also measured.

We can lengthen a list of the conversational alternatives 'yes' and 'no' with a middle alternative (i.e., 'maybe'), without increasing the chance of mismatch answers. If longer lists are preferred, then it is necessary to adequately instruct interviewers how to present response alternatives and probe for adequate answers.

However, there is no straightforward solution how such longer lists should be presented to respondents. In the Television survey data, attitude direction and extremity were asked in a two-step procedure. First, the direction of the opinion was asked, and next the intensity of the opinion was asked. This procedure yielded a high percentage of mismatch answers (i.e., in 55% of the Q-A sequences). Moreover, interviewers appeared to skip the second step of the procedure (i.e., avoided probing for intensity of the opinion). This may be due to laziness of interviewers, or to the possibility that interviewers infer that respondents hardly understand the difference between 'strongly' and 'just' (dis)agree.

8.2.3 Effects of memory and type of response alternatives

It may also be useful to gather more knowledge about the effects of the conversational character of response alternatives and memory. We do not know if there is a difference in the effort respondents need to invest in remembering the exact wording of conversational or formal alternatives. Conversational alternatives may be easily available, but can also create confusion. Holbrook's (2000) study showed that unconventional word orders for response alternatives disrupt cognitive processing, thus we can also argue that unconventional (formal) words as response alternatives disrupt cognitive processing, which may have the result that it is more difficult to remember such alternatives. However, formal words may also be easier to remember because they may be more salient.

8.2.4 Order of questions and types of alternatives

Further research could also focus on the order of different types of alternatives and questions in the questionnaire. The experimental survey and the non-experimental ESS-study, showed that frequent switches in the use of different response alternatives increase the number of mismatch answers occurring in an interview. In the experimental survey, we found that questions with conversational alternatives that were immediately preceded by questions with formal alternatives yielded more mismatch answers than conversational alternatives that were not preceded by formal alternatives. This effect showed that respondents, once they had learned to use the formal alternatives, kept using those alternatives for subsequent questions. We did not find an opposite question-order effect: questions with formal alternatives that were preceded by questions with conversational alternatives did not yield more mismatch answers than the same questions that were not preceded by questions with conversational alternatives. However, question-order effects were mainly avoided by means of buffer questions with a totally different response format (i.e., implicit alternative questions that clearly imply a different response format such as 'how many days a week do you watch television'). Thus, when the use of different response alternatives may be unavoidable in a survey, it is advisable not to ask questions with formal alternatives directly before questions with conversational alternatives.

8.2.5 Effects of interviewers

In this thesis we did not study systematic differences in interviewer behaviors. It is possible that interviewers increase the chance of mismatch answers. For example, interviewers may differ in their politeness, their ability to recognize problems in misunderstanding, and their commitment to adhere to standardization rules. In this way, interviewers may differ with respect to the conversational character of the interview they evoke. They may motivate respondents to elaborate, and make the interview a pleasant experience. This will confirm respondents' idea that they are being cooperative when they provide mismatch answers.

8.2.6 Conversion of mismatch answers into adequate answers

Interviewers obviously play an important role in the solution of mismatch answers. It is certainly useful to learn interviewers to recognize mismatch answers, and to inform them about the behaviors they are surely not allowed to perform (e.g., probe suggestively or score responses based upon inferences of mismatch answers).

As the different types of mismatch answers have different causes, they also require different actions of the interviewers. However, it may be difficult to immediately recognize the different types of mismatch answers. Conversational mismatch answers occur most frequently and therefore it is most useful to instruct interviewers to repeat the alternatives non-directively whenever a mismatch answer is given. According to strictly standardized rules all alternatives must be repeated (Fowler and Mangione 1990), but from a more moderate approach we would recommend to instruct interviewers that it is not necessary to

mention all response alternatives, but at least two alternatives that are warranted by the respondent's mismatch answer (and never only one alternative).

In addition, interviewers can be trained to pay special attention to cues that may indicate task uncertainty or state uncertainty, to adapt their strategy in helping respondents to format their answer adequately. Such cues are for example, verbal expressions of uncertainty or long response latencies.

Task mismatch answers may be most difficult to be solved adequately. A task mismatch is an attempt of respondents to answer, whereas they are not certain of a correct understanding of the question. In order to solve the problem of task uncertainty, interviewers have to clarify question meaning. When interviewers have not been trained adequately to do this, they may cause problematic deviations by giving inadequate clarifications or suggesting an answer. How important it is that interviewers deal with problems in understanding in a standardized way remains to be seen. Although variation in interviewer behavior may arise when they are given the freedom to clarify questions in their own words, we would agree that it is more important that interviewers know the questions and concepts, than that they literally keep to the script.

Finally, interviewers can help solve cognitive mismatch answers by stimulating the respondent to retrieve more information from memory. Respondents are usually well aware of the fact that their mismatch answer creates a problem for the interviewer, and that they themselves are the only ones who have the information available to answer the question. Thus, when interviewers try to solve the cognitive mismatch answers by means of a suggestive probe, respondents will not necessarily confirm this suggestion.

8.2.7 Validity of conversational responses

We do not know if conversational responses, that are adequate when conversational response alternatives are used, also yield more valid and reliable data. One of the possible consequences of listing conversational responses is that, while allowing respondents to use convenient words, respondents may view their task as less important, and as a result process questions less thoroughly. They may automatically give a conversational response, which accidentally happens to be an adequate response, as the alternatives are conversational too. Thus, the interviewer does not need to probe for adequately formatted answers. In this way they are hardly reminded of the formal character of the survey. In case of formal alternatives, however, they are reminded of the formal character of the task

However, in our opinion conversational alternatives generally yield responses of higher quality. Mismatch answers do not necessarily yield inaccurate responses by itself: it depends on what the interviewer does to solve them. Conversational response alternatives generate less mismatch answers, and consequently less need for probing, whereas formal response alternatives increase the chance of mismatch answers, increasing the need for probing. As long as interviewers frequently probe suggestively, or even infer from the mismatch answer what score to fill in without any probing at all, mismatch answer will eventually lead to less valid scores. As Houtkoop-Steenstra states it "Because research methodology favors

standardization of the interviewing process, designing questions that may lead to the need for probing […] should be avoided" (Houtkoop-Steenstra, 2000, p. 119).

Although the relation between validity and the occurrence of mismatch answers has been studied before, it will be most interesting to compare the reliability and validity of answers to questions with conversational and formal alternatives in forthcoming studies.

8.3 Summary and recommendations

The results presented in this thesis showed that respondents are usually responsible for the first problems occurring in Q-A sequences, and this mostly is a mismatch answer. Because of these mismatch answers, interviewers also behave problematically (i.e., probe suggestively or infer responses without verifying them), which may have negative consequences for the quality of the response obtained. Furthermore, even if interviewers know how to respond adequately to convert mismatch answers into adequate ones, this conversion lengthens the interaction, and thus increases the costs of survey interviews. Therefore, it is important to reduce the number of mismatch answers that occur in survey interviews. We have a few recommendations that can be implemented to reduce the chance of mismatch answers, or to reduce the negative consequences of mismatch answers:

- Use response alternatives that adequately match the question. Thus when a question is worded as a yes-no question, the response alternatives must be just 'yes' and 'no' (and 'don't know' or refuses to answer' if applicable).
- Use response alternatives that are adapted to the conversational style of responding. Such response alternatives consist of words that are most commonly used in ordinary conversations.
- For assertions 'yes' and 'no' should be used as alternatives, or assertions should have adapted wording such as "To what extent do you find that…" Mismatch answers are very likely to occur with assertions in the way they are usually worded in regular surveys. Respondents treat assertions as yes-no questions, i.e., they typically reply with 'yes' or 'no' and such answers do not fit with the five-point Likert-type scale.
- Do not switch more than absolutely necessary between different response formats in a questionnaire (in order to prevent confusion), or use 'buffer questions', with a completely different response format (preferably open-ended questions with implicit alternatives).
- If questions with formal alternatives are unavoidable, do not ask them before questions with conversational alternatives.
- Use show cards to present response alternatives in face-to-face interviews
- Use formal question wording, to remind respondents of the formal character of the survey, which may stimulate them to give more precisely formatted answers.
- Learn interviewers to recognize mismatch answers, and train them in the use of adequate reactions (repeat the response alternatives, and never offer only one alternative).

References

AAPOR. (2004) "Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 3rd edition." Lenexa, Kansas: The American Association for Public Opinion Research.

Bakeman, R., and J.M. Gottman. (1997) *Observing interaction: An introduction to sequential analysis*. Cambridge: University Press.

Bates, N., and C. Good. (1996) "An evaluation of the 1995 Test Census Integrated Coverage Measurement (ICM) Interview: Results from Behavior Coding." Paper presented at *Annual meeting of the American Statistical Association* Chicago

Beatty, P. (1995) "Understanding the Standardized/Non-Standardized Interviewing Controversy." *Journal of Official Statistics* 11 Pp. 147-160.

Beatty, P. (2004) "The Dynamics of Cognitive Interviewing." in *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J. Rothgeb, M. P. Couper, J.T. Lessler, E. Martin, J. Martin and E. Singer. New York: Wiley.

Belli, R. F., N. Schwarz, E. Singer, and J. Talarico. (2000) "Decomposition can harm the accurarcy of behavioural frequency reports." *Applied cognitive psychology* 14 Pp. 295-308.

Belli, R.F., E.H. Lee, F.P. Stafford, and C.-H. Chou. (2004) "Calendar and Question-List Survey Methods: Association Between Interviewer Behaviors and Data Quality." *Journal of Official Statistics* 20 Pp. 185 - 218.

Belli, R.F., and J.M. Lepkowski. (1996) "Behavior of Survey Actors and the Accuracy of Response." *Health Survey Research Methods: Conference Proceedings* DHMS Publication No. (PHS)96-1013 Pp. 69-74.

Belli, R.F., J.M. Lepkowski, and M.U. Kabeto. (2001) "The Respective Roles of Cognitive Processing Difficulty and Conversational Rapport on the Accuracy of Retrospective Reports of Doctor's Office Visits." in *Seventh Conference on Health Survey Research Methods*, edited by M.L. Cynamon and R.A. Kulka. Hyattsville, MD: U.S. Government Printing Office, (DHHS Publication No. (PHS) 01-1013).

Belson, W.A. (1981) *The Design and Understanding of Survey Questions*. Aldershot: Gower.

Bernts, T. (1991) *Leven zonder zorg. Oordelen over risico's, rechtvaardigheid en solidariteit in de gezondheidszorg.* Amsterdam: Swets en Zeitlinger.

Biemer, P. (1988) "Measuring data quality." in *Telephone Survey Methodology*, edited by R. M. Groves, P. Biemer, L. Lyberg, W.L. Massey and J. Waksberg. New York: John Wiley.

Biemer, P., D. Herget, J. Morton, and G. Willis. (2000) "The feasibility of monitoring field interview performance using computer audio recorded interviewing (CARI)." in *Proceedings of the section of survey research methods*. Alexandria, V.A.: American Statistical Association.

Blair, E. (1978) *Nonprogrammed Speech Behaviors in a Household Survey*: unpublished doctoral dissertation, University of Illinois, Department of Business Administration.

Blair, E. (1980) "Using Practice Interviews to Predict Interviewer Behaviors." *Public Opinion Quarterly* 44 Pp. 257-260.

Blixt, S., and J. Dykema. (1995) "Before the pretest: Question Development Strategies." in *Proceedings of the section of survey research methods*. Alexandria: V.A.: American Statistical Association.

Bradburn, N.M., and S. Sudman. (1979) *Improving Interview Method and Questionnaire design; Response Effects to Threatening Questions in Survey Research.* San Francisco: Jossey-Bass.

Brennan, S.E. (forthcoming) "The Vocabulary Problem in Spoken Dialogue Systems." in *Automated Spoken Dialog Systems*, edited by S. Luperfoy. Cambridge, MA: MIT Press.

Brennan, S.E., and H.H. Clark. (1996) "Conceptual Pacts and Lexical Choice in Conversation." *Journal of Experimental Psychology: Learning, Memory and Cognition* 22 Pp. 1482-1492.

Brenner, M. (1982) "Response-Effects of "Role-Restricted" Characteristics of the Interviewer." in *Response Behaviour in the Survey-Interview*, edited by W. Dijkstra and J. Van der Zouwen. London: Academic Press.

Brick, J.M., M.A. Collins, M.J. Nolin, E. Davies, and M.L. Feibus. (1997a) "Design, Data Collection, Monitoring, Interview Administration Time, and Data Editing in the 1993 National Household Education Survey." Washington, D.C.: U.S. Department of Education. National Center for Education Statistics.

Brick, J.M., E. Tubbs, M.A. Collins, M.J. Nolin, D. Cantor, K. Levin, and Y Carnes. (1997b) "Telephone Coverage Bias and Recorded Interviews in the 1993 National Household Education Survey." Washington, D.C.: National Center for Education Research.

Brook, R.H., J.E. Ware jr., A. Davies-Avery, A.L Stewart, C.A. Donald, W.H Rogers, K.N Williams, and S.A. Johnston. (1979) "Overview of adult health status measures fielded in Rand's Health Insurance Study." *Med Care* 17.

Brown, P., and S. Levinson. (1987) *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Burgess, M.J., and D. Patton. (1993) "Coding of respondent behaviour by interviewers to test questionnaire wording." *Proceedings of the ASA Section of Survey Research Methods* Pp. 392-397.

Cahalan, M., S. Mitchell, L. Gray, S. Chen, and J. Tsapogas. (1994) "Recorded interview behavior coding study: national survey of recent college graduates." *Proceedings of the ASA Section on Survey Research Methods.*

Campanelli, P. (1997) "Testing Survey Questions: New Directions in Cognitive Interviewing." *Bulletin de Méthodologie Sociologique* 55 Pp. 5-17.

Cannell, C.F., F.J. Fowler, and K.H. Marquis. (1968) "The influence of interviewer and respondent psychological and behavioral variables on the reporting of household interviews." *Vital and Health Statistics, Series 2, No. 26.*

Cannell, C.F., and R.L. Kahn. (1953) "The collection of data by interviewing." in *Research Methods in the Behavioral Sciences.*, edited by L. Festinger and D. Kats. New York: The Dryden Press.

Cannell, C.F., and R.L. Kahn. (1968) "Interviewing." in *The Handbook of Social Psychology, Vol. 2*, edited by G. Lindzey and E. Aronson. Reading, Mass.: Addison-Wesley.

Cannell, C.F., S.A. Lawson, and D.L. Hausser. (1975) "A Technique for Evaluating Interviewer Performance: A Manual for Coding and Analyzing Interviewer Behavior from Tape Recordings of Household Interviews." Ann Arbor, MI: Survey Research Center of the Institute for Social Research, The University of Michigan,.

Cannell, C.F., P.V. Miller, and L. Oksenberg. (1981) "Research on Interviewing Techniques." in *Sociological Methodology 1981*, edited by S. Leinhardt. San Francisco: Jossey-Bass.

Cannell, C.F., and L. Oksenberg. (1988) "Observation of Behavior in Telephone Interviews." in *Telephone Survey Methodology*, edited by R. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls II and J. Waksberg. New York: Wiley.

Cannell, C.F., L. Oksenberg, and J.M. Converse. (1977) "Striving for response accuracy: Experiments in New Interviewing Techniques." *Journal of Marketing Research* XIV Pp. 306-315.

Carton, A. (1999) "Een Interviewnetwerk: Uitwerking van een Evaluatieprocedure voor Interviewers." Leuven:: Proefschrift Faculteit Sociale Wetenschappen.

Churchill, L. (1978) *Questioning Strategies in Sociolinguistics*. Rowley, Mass.: Newbury House.

Cicourel, A. (1982) "Interviews, Surveys, and the Problem of Ecological Validity." *The American Sociologist* 17 Pp. 11-20.

Clark, H.H. (1985) "Language Use and Language Users." in *Handbook of Social Psychology*, edited by G. Lindzey and E. Aronson. New York: Random House.

Clark, H.H., and M. F. Schober. (1992) "Asking questions and influencing answers." in *Questions about questions.*, edited by J.M. Tanur. New York: Russell Sage Foundation.

Conrad, F. G., and M. F. Schober. (2000) "Clarifying question meaning in a household telephone survey." *Public Opinion Quarterly* 64 Pp. 1-28.

Couper, M.P., L. Holland, and R.M. Groves. (1992) "Developing Systematic Procedures for Monitoring in a Centralized Telephone Facility." *Journal of Official Statistics* 8 Pp. 63-76.

De Waal, E., K. Schonbach, and E. Lauf. (2004) " Online Newspapers: A Substitute for Print Newspapers and Other Information Channels?" Paper presented at *6th World Media Economics Conference* HEC Montreal, Canada

DeMaio, T.J., N. A. Mathiowetz, J. Rothgeb, M.E. Beach, and S. Durant. (1993) "Protocol for Pretesting Demographic Surveys at the Census Bureau." *Proceedings of the Section on Survey Research Methods*.

Dijkstra, W. (1983) *Beïnvloeding van antwoorden in survey-interviews.* Academisch proefschrift Vrije Universiteit. Utrecht: Elinkwijk.

Dijkstra, W. (1999) "A New Method for Studying Verbal Interactions in Survey Interviews." *Journal of Official Statistics* 15 Pp. 67-85.

Dijkstra, W. (2002) "Transcribing, Coding, and Analyzing Verbal Interactions in Survey Interviews." in *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview.*, edited by D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer and J. Van der Zouwen. New York: Wiley.

Dijkstra, W., and Y.P. Ongena. (forthcoming) "Question-Answer Sequences in Survey-Interviews." *Quality and quantity*.

Dijkstra, W., L. Van der Veen, and J. Van der Zouwen. (1985) "A Field Experiment on Interviewer-Respondent Interaction." in *The Research Interview; Uses and Approaches*, edited by M. Brenner, J. Brown and D. Canter. London etc.: Academic Press.

Dijkstra, W., and J. Van der Zouwen. (1982) "Response Behaviour in the Survey-Interview." London: Academic Press.

Dijkstra, W., and J. van der Zouwen. (1988) "Types of inadequate interviewer behaviour in survey-interviews." in *Sociometric research. Volume 1: Data collection and scaling*, edited by W. E. Saris and I. N. Gallhofer. London: MacMillan.

Dorsey, B.L., O.N. Rosemery, and S.C. Hayes. (1986) "The effects of code complexity and of behavioral frequency on observer accuracy and interobserver agreement." *Behavioral assessment* 8 Pp. 349-363.

Draisma, S. , and W. Dijkstra. (2004) "Response Latency and (Para)linguistic Expressions as Indicators of Response Error." in *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J Rothgeb, M. P. Couper, J.T. Lessler, E. Martin, J. Martin and E. Singer. New York: Wiley.

Draisma, S., W. Dijkstra, and Y.P. Ongena. (forthcoming) "Qualified answers and other doubt expressions as indicators of cognitive problems in a health survey." *Proceedings of the Survey Research Methods Section*.

Dykema, J., J.M. Lepkowski, and S. Blixt. (1997) "The Effect of Interviewer and Respondent Behavior on Data Quality: Analysis of Interaction Coding in a Validation Study." in *Survey Measurement and Process Quality.*, edited by L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz and D. Trewin. New York: John Wiley & Sons Inc.

Edwards, W.S., S. Fry, E. Zahnd, N Lordi, and G. Willis. (2004) "Behavior coding across multiple languages: The 2003 California Health Interview Survey as a case study." *Proceedings of the Joint Statistical Meetings, Survey Research Methods Section.*

Edwards, W.S., V. Narayanan, S. Fry, J.A. Catania, and M. Pollack. (2002) "A Comparison of two behavior coding systems for pretesting questionnaires." *Proceedings of the Joint Statistical Meetings, Survey Research Methods Section* Pp. 889-892.

Elchardus, M., C. Tresignie, and A. Derks. (undated) "Project: het draagvlak van de solidariteit. Deelrapport 3: Levensstijl- en levensloopaansprakelijkheid." Vakgroep Sociologie, Onderzoeksgroep TOR, Vrije Universiteit Brussel.

Esposito, J.L., J. Rothgeb, A.E. Polivka, J. Hess, and P.C. Campanelli. (1992) "Methodologies for Evaluating Survey Questions: Some Lessons from the Redesign of the Current Population Survey." Paper presented at *International Conference on Social Science Methodology,* Trento

ESS. (2005) "European Social Survey." http://www.europeansocialsurvey.org/.

Fowler, F. J. (1992) "How unclear terms affect survey data." *Public Opinion Quarterly* 56 Pp. 218-231.

Fowler, F. J. (2002) *Survey Research Methods*. Thousand Oaks: Sage.

Fowler, F.J., and Th. W. Mangione. (1990) *Standardized Survey Interviewing; Minimizing Interviewer-Related Error*. Newbury Park, CA: Sage.

Gallagher, P. M., F. J. Fowler, and A. Roman. (2004) "Training Elderly Respondents: Does it help?" Paper presented at *59th Annual Conference of the American Association for Public Opinion Research* Phoenix, Arizona

Goffman, E. (1967) *Interaction ritual: Essays on face-to-face behavior*. Garden City, NY: Anchor & Doubleday.

Grice, H.P. (1975) "Logic and Conversation." in *Syntax and Semantics 3: Speech Acts*, edited by P. Cole and J.L. Morgan. New York: Academic Press.

Gustavson-Miller, L.A., D.J. Herrman, and M.C. Puskar. (1991) "The Effects of Question Reading on Respondent Verbal Behavior." Washington D.C.: Bureau of Labor Statistics.

Hansen, S. E., and M. P. Couper. (2004) "Usability testing to Evaluate Computer-Assisted Instruments." in *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J. Rothgeb, M. P. Couper, J.T. Lessler, E. Martin, J. Martin and E. Singer. New York: Wiley.

Heritage, J. (1984) *Garfinkel and Ethnomethodology*. Cambridge: Polity.

Heritage, J. (2002) "Ad Hoc Inquiries: Two Preferences in the Design of Routine Questions in an Open Context." in *Standardization and Tacit Knowledge: Interaction and*

Practice in the Survey Interview, edited by D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer and J. Van der Zouwen. New York: Wiley.

Hess, J., J. Rothgeb, and A.L. Zukerberg. (1997) "Pretest Evaluation Report. Survey of Program Dynamics." Center for Survey Methods Research, U.S. Census Bureau.

Hill, D.H., and J.M. Lepkowski. (1996) "Behavioral Contagion in the Health Field Survey." Paper presented at *Health Survey Research Methods: Conference Proceedings*

Holbrook, A.L. , J. A. Krosnick, R.T. Carson, and R.C. Mitchell. (2000) "Violating Conversational Conventions Disrupts Cognitive processing of attitude questions." *Journal of Experimental Social Psychlogy* 36 Pp. 465-494.

Houtkoop-Steenstra, H. (1994) "Het genereren van passende antwoorden in survey interviews." Paper presented at *Zesde Sociaal-wetenschappelijke Studiedagen 1994, 7 & 8 april 1994, Amsterdam*

Houtkoop-Steenstra, H. (2000) *Interaction in the standardized survey interview: the living questionnaire*. Cambridge: Cambridge University Press.

Houtkoop-Steenstra, H. (2002) "Questioning turn format and turn-taking problems in standardized interviews." in *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, edited by D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer and J. Van der Zouwen. New York: Wiley.

Hughes, K.A. (2004) "Comparing Pretesting Methods: Cognitive Interviews, Respondent Debriefing, and Behavior Coding." Statistica; Research Division U.S. Bureau of the Census.

Hyman, H., W.J. Cobb, J. Feldman, C.W. Hart, and C. Stember. (1954) *Interviewing in Social Research*. Chicago: University of Chicago Press.

Jefferson, G. (1983) "Notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation." in *TILL-paper no. 42. Tilburg papers in Language and literature*: Katholieke Universiteit Brabant.

Kalton, G., and H. Schuman. (1982) "The Effect of the Question on Survey Responses: A Review." *Journal of the Royal Statistical Society* 145 Pp. 42-57.

Kempen, KG.I.J.M., E.I. Brilman, J.W. Heyink, and J. Ormel. (1995) "Het meten van de algemene gezondheidstoestand met de MOS Short-Form General Health Survey (SF-20): een handleiding." Groningen: Noordelijk Centrum voor Gezondheidsvraagstukken, Rijksuniversiteit Groningen.

König-Zahn, C., J.W. Furer, and B. Tax. (1993) *Algemene gezondheid*. Assen: Van Gorcum.

Kriegsman, D.M.W., J.T.M. Van Eijk, and D.J.H. Deeg. (1995) "Psychometrische eigenschappen van de Nederlandse versie van de RAND General Health Perception Questionnaire: de Vragenlijst Algemene Gezondheidsbeleving (VAGB) [Psychometric properties of the RAND General Health Perception Questionnaire in The Netherlands]." *Tijdschrift Sociale Gezondheidszorg* 73 Pp. 390-398.

Krosnick, J. A. (1999) "Survey research." *Annual Review of Psychology* 50 Pp. 537-567.

Krosnick, J.A., and R.P. Abelson. (1992) "The Case for Measuring Attitude Strength in Surveys." in *Questions about Questions*, edited by J. Tanur. New York: Russell Sage Foundation.

Landis, J.R., and G.G. Koch. (1977) "The measurement of observer agreement for categorical data." *Biometrics* 45 Pp. 255-268.

Leeuw, E. de, and W. de Heer. (2002) "Trends in Household Survey Nonresponse: A Longitudinal and International Comparison." in *Survey Nonresponse*, edited by R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little. New York: Wiley.

Lepkowski, J.M., M. P. Couper, S. E. Hansen, W. Landers, K.A. McGonagle, and J. Schlegel. (1998) "CAPI Instrument Evaluation: Behavior Coding, Trace files, and

Usability Testing." in *Proceedings of the section of Survey Research Methods*. Alexandria, V.A.: American Statistical Association.

Lepkowski, J.M., V. Siu, and J. Fisher. (2000) "Event History Analysis of Interviewer and Respondent Survey Behavior." in *Developments in Survey Methodology*, edited by A. Ferligoj and A. Mrvar. Ljubljana: FDV.

Loosveldt, G. (1985) "De effecten van een interviewtraining op de kwaliteit van gegevens bekomen via het survey-interview." Leuven: Proefschrift Katholieke Universiteit Leuven.

Loosveldt, G. (1994) "The Profile of the Difficult-to-Interview Respondent." Paper presented at *13th World Congress of Sociology* Bielefeld

Loosveldt, G. (1995) "Interviewer-respondent interaction analysis as a diagnostic and validating instrument." Paper presented at *The International Conference on Survey Measurement and Process Quality* Bristol

Loosveldt, G. (1997) "Interaction Characteristics of the Difficult to interview respondent." *International Journal of Public Opinion Research* 9 Pp. 386-394.

Mallinson, S. (2002) "Listening to respondents: a qualitative assessment of the Short-Form 36 Health Status Questionnaire." *Social Science & Medicine* Pp. 11-21.

Marquis, K.H., and C. F. Cannell. (1969) *A Study of Interviewer-Respondent Interaction in the Urban Employment Survey*. Ann Arbor: Mich. Survey Research Center. University of Michigan.

Mathiowetz, N. A. (1999) "Respondent uncertainty as indicator of response quality." *International Journal of Public Opinion Research* 11 Pp. 289-296.

Mathiowetz, N. A., and C. F. Cannell. (1980) "Coding Interviewer Behavior as a Method of Evaluating Performance." in *Proceedings of the section of survey research methods*. Alexandria, V.A.: American Statistical Association.

Maynard, D.W., and N.C. Schaeffer. (2002) "Standardization and Its Discontents." in *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, edited by D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer and J. Van der Zouwen. New York: Wiley.

Mazeland, H. (2003) *Inleiding in de Conversatie Analyse*. Bussum: Coutinho.

Means, B., G.E. Swan, J.B. Jobe, and J.L. Esposito. (1991) "An Alternative Approach to Obtaining Personal History Data." in *Measurement Errors in Surveys*, edited by P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman. New York: Wiley.

Molenaar, N.J., and J.H. Smit. (1996) "Asking and Answering Yes/No-Questions in Survey Interviews: A Conversational Approach." *Quality and Quantity* 30 Pp. 115-136.

Moore, R.J. (2004) "Managing troubles in answering survey questions: respondents' uses of projective reporting." *Social Psychology Quarterly* 67 Pp. 50-69.

Moore, R.J., and D.W. Maynard. (2002) "Achieving Understanding in the Standardized Survey Interview: Repair Sequences." in *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, edited by D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer and J. Van der Zouwen. New York: Wiley.

Morton-Williams, J. (1979) "The use of 'verbal interaction coding' for evaluating a questionnaire." *Quality and Quantity* 13 Pp. 59-75.

Oksenberg, L., C. F. Cannell, and S. Blixt. (1996) "Analysis of Interviewer and Respondent Behavior in the Household Survey. National Medical Expenditure Survey Methods 7, Agency for Health Care Policy and Research." Rockville, MD: Public Health Service, U.S. Department of Health and Human Services.

Oksenberg, L., C. F. Cannell, and G. Kalton. (1991) "New Strategies for Pretesting Survey Questions." *Journal of Official Statistics* 7 Pp. 349-365.

Ongena, Y.P., and W. Dijkstra. (forthcoming) "Methods of behavior coding of survey interviews." *Journal of Official Statistics*.

Ongena, Y.P., W. Dijkstra, and S. Draisma. (2004) "Conversational and formal questions in Survey Interviews." Paper presented at *The Sixth International Conference on Logic and Methodology RC-33.* Amsterdam

Oostdijk, N., W. Goedertier, F. van Eynde, L. Boves, J. Martens, M. Moortgat, and H. Baayen. (2002) "Experiences from the Spoken Dutch Corpus Project." Paper presented at *Proceedings of the Third International Conference on Language Resources and Evaluation*

Payne, S.L. (1951) *The Art of Asking Questions*. Princeton: Princeton University Press.

Petty, R.E., and J.T. Cacioppo. (1981) *Attitudes and Persuasion: Classic and Contemporary Approaches.* Dubuque, Iowa: Wm. C. Brown Publishers.

Petty, R.E., G.A. Rennier, and J.T. Cacioppo. (1987) "Assertion Versus Interrogation Format in Opinion Surveys: Questions Enhance Thoughtful Responding." *Public Opinion Quarterly* 51 Pp. 481-494.

Presser, S., and J. Blair. (1994) "Survey Pretesting: Do different Methods Produce Different Results?" *Sociological Methodology* 24 Pp. 73-194.

Prüfer, P., and M. Rexroth. (1985) "Zur Anwendung der Interaction-Coding-Technik." *ZUMA-Nachrichten* 17 Pp. 2-49.

RVD. (2003) "Belevingsmonitor Rijksoverheid, juni 2003." Den Haag: Rijksvoorlichtingsdienst/Publiek en Communicatie.

Sacks, H., E. Schegloff, and G. Jefferson. (1974) "A simplest systematics for the organization of turn-taking in conversation." *Language* 50 Pp. 696-735.

Sander, J.E., F.G. Conrad, P.A. Mullin, and D.J. Herrmann. (1992) "Cognitive modelling of the survey interview." *Proceedings of the Joint Statistical Meetings, Survey Research Methods Section.* Pp. 818-823.

Schaeffer, N. C., and S. Presser. (2003) "The Science of Asking Questions." *Annual Review of Sociology* 29 Pp. 65-88.

Schaeffer, N.C. (1991) "Conversation with a Purpose - Or Conversation? Interaction in the Standardized Interview." in *Measurement Errors in Surveys*, edited by P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman. New York: Wiley.

Schaeffer, N.C. (2002) "Conversation with a Purpose-or Conversation? Interaction in the Standardized Interview." in *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, edited by D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer and J. Van der Zouwen. New York: Wiley.

Schaeffer, N.C., and J. Dykema. (2004) "Improving the Clarity of Closely Related Concepts: Distinguishing Legal and Physical Custody of Children." in *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J. Rothgeb, M. P. Couper, J.T. Lessler, E. Martin, J. Martin and E. Singer. New York: Wiley.

Schaeffer, N.C., and D. W. Maynard. (1996) "From Paradigm to Prototype and Back again: Interactive Aspects of Cognitive Processing in Standardized Survey Interviews." in *Answering Questions. Methodology for Determining Cognitive and Communicative Processes in Survey Research*, edited by N. Schwarz and S. Sudman. San Francisco: Jossey-Bass.

Schaeffer, N.C., and D. W. Maynard. (2002) "Occasions for Intervention: Interactional Resources for Comprehension in Standardized Survey Interviews." in *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, edited by D.

W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer and J. Van der Zouwen. New York: John Wiley & Sons, Inc.

Schaeffer, N.C., and E. Thomson. (1992) "The Discovery of Grounded Uncertainty; Developing Standardized Questions about Strength of Fertility Motivations." in *Sociological Methodology 1992, Volume 22*, edited by P.V. Marsden. Oxford: Basil Blackwell.

Schegloff, E., and H. Sacks. (1973) "Opening up Closings." *Semiotica* 8 Pp. 289-327.

Schober, M. F. (1999) "Making Sense of Questions: An Interactional Approach." in *Cognition and Survey Research*, edited by M.G. Sirken, D.J. Herrman, S. Schechter, N. Schwarz, J.M. Tanur and R. Tourangeau: John Wiley & Sons Inc.

Schober, M. F., and F.G. Conrad. (1997) "Does conversational interviewing reduce survey measurement error?" *Public Opinion Quarterly* 61 Pp. 576-602.

Schober, M. F., and F.G. Conrad. (2002) "A Collaborative View of Standardized Survey Interviews." in *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, edited by D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer and J. Van der Zouwen. New York: Wiley.

Schuman, H., and S. Presser. (1981) *Questions and Answers in Attitude Surveys; Experiments on Question Form, Wording, and Context*. New York: Academic Press.

Schwarz, N. (1995) "Questionnaires: The Survey Interview and the Logic of Conversation." *International Statistical Review* Pp. 153-169.

Schwarz, N. (1996) *Cognition and Communication: Judgemental Biases, Research Methods, and the Logic of Conversation*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Schwarz, N., and D. Oyserman. (2001) "Asking questions about behavior: Cognition, communication, and questionnaire construction." *American Journal of Evaluation* 22 Pp. 127-160.

Shepherd, J., and C. Vincent. (1991) "Interviewer-Respondent Interactions in CATI interviews." in *Proceedings of the annual research conference*: Bureau of the Census.

Slugoski, B.R., and D.J. Hilton. (2001) "Conversation." in *The New Handbook of Language and Social Psychology*, edited by W.P. Robinson and H. Giles. Chichester: John Wiley & Sons.

Smit, E.G., and P.C. Neyens. (2000) "Segmentation based on Affinity for Advertising." *Journal of Advertising Research* 40 Pp. 35-43.

Smit, J. H. (1995) *Suggestieve vragen in survey-interviews. Vóórkomen, oorzaken en gevolgen*. Amsterdam: Academisch proefschrift, Vrije Universiteit.

Smit, J.H., W. Dijkstra, and J. Van der Zouwen. (1997) "Suggestive Interviewer Behaviour in Surveys: An Experimental Study." *Journal of Official Statistics* 13 Pp. 19 - 28.

Snijders, T.A.B., and R.J. Bosker. (1999) *Multilevel Analysis. An introduction to basic and advanced multilevel modeling*. London: Sage publications.

Snijkers, G.J.M.E. (2002) *Cognitive laboratory experiences. On pre-testing computerised questionnaires and data quality*. Utrecht: Proefschrift Universiteit Utrecht.

Stanley, J.S. (1996) "Standardizing Interviewer Behavior based on the Results of Behavior Coding." *Proceedings of the Joint Statistical Meetings, Survey Research Methods Section.*

Stax, H.-P. (2004) "Paths to precision: probing turn format and turn-taking problems in standardized interviews." *Discourse Studies* 6 Pp. 77-94.

Suchman, L., and B. Jordan. (1990) "Interactional troubles in face-to-face survey interviews." *Journal of the American Statistical Association* (85) Pp. 232-253.

Sudman, S., and N.M. Bradburn. (1974) *Response Effects in Surveys.* Chicago: Aldine Publishing Company.

Sudman, S., N.M. Bradburn, and N. Schwarz. (1996) *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.

Sykes, W., and M. Collins. (1992) "Anatomy of the Survey Interview." *Journal of Official Statistics* 8 Pp. 277-291.

Sykes, W., and J. Morton-Williams. (1987) "Evaluating Survey Questions." *Journal of Official Statistics* 2 Pp. 191-207.

Tarnai, J., and D.L. Moore. (2004) "Methods for Testing and Evaluating Computer-Assisted Questionnaires." in *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J. Rothgeb, M. P. Couper, J.T. Lessler, E. Martin, J. Martin and E. Singer. New York: Wiley.

Tourangeau, R., L. C. Rips, and Rasinski. (2000) *The psychology of survey response*. Cambridge: Cambridge University Press.

Van de Berg, J, R. Jansen, and H. Haveman. (1986) "Solidariteitsvoorkeuren ten aanzien van ziektekostenverzekering 1985." *Maandbericht gezondheidsstatistiek* 5 Pp. 5-16.

Van der Zouwen, J. (2001) "Een typologie van methoden voor het beoordelen van vragen(lijsten)." *IOPS-course Introduction*.

Van der Zouwen, J., and W. Dijkstra. (1995) "Trivial and non-trivial question-answer sequences; types, determinants and effects on data quality." in *Proceedings of the international Conference on Survey Measurement and Process Quality*. Bristol. (UK).

Van der Zouwen, J., and W. Dijkstra. (1998) "Het vraaggesprek onderzocht. Wat zegt het verloop van de interactie in survey-interviews over de kwaliteit van de vraagformulering?" *Sociologische gids* 45 Pp. 387-403.

Van der Zouwen, J., W. Dijkstra, and J. H. Smit. (1991) "Studying Respondent-Interviewer Interaction: The Relationship Between Interviewing Style, Interviewer Behavior, and Response Behavior." in *Measurement Errors in Surveys.*, edited by P. Biemer, R. Groves, L. Lyberg, N. A. Mathiowetz and S. Sudman. New York: Wiley.

Van der Zouwen, J., and J. H. Smit. (2004) "Evaluating Survey Questions by Analyzing Patterns of Behavior Codes and transcripts of Question-Answer Sequences: A Diagnostic Approach." in *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J. Rothgeb, M. P. Couper, J.T. Lessler, E. Martin, J. Martin and E. Singer. New York: Wiley.

Viterna, J.S., and D.W. Maynard. (2002) "How Uniform Is Standardization? Variation Within and Across Survey Research Centers Regarding Protocols for Interviewing." in *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, edited by D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer and J. v.d. Zouwen. New York: Wiley.

Watson, D. (1982) "The Actor and the Observer: How Are their Perceptions of Causality Divergent?" *Psychological Bulletin*.

Weiss, C.H. (1968) "Validity of welfare mothers' interview responses." *Public Opinion Quarterly* 32 Pp. 622-633.

Weiss, C.H. (1970) "Interaction in the research interview; the effects of rapport on response." *Proceedings of the Social Statistics Section, American Statistical Association* Pp 17-20.

Wilcox, K. (1963) "Comparison of Three Methods for Collection of Morbidity Data by Household Survey." in *The University of Michigan, School of Public Health*.

Willis, G. (2005) *Cognitive Interviewing. A Tool for Improving Questionnaire Design.* Thousand Oaks: Sage.

Willis, G., and J.T. Lessler. (1999) "Question appraisal system." Rockville, M.D.: Research Triangle Institute.

Willis, G., S. Schechter, and K. Whitaker. (1999) "A comparison of cognitive interviewing, expert review and behavior coding: What do they tell us?" *Proceedings of the Joint Statistical Meetings, Survey Research Methods Section.* Pp. 28-37.

Appendices

Appendix 4-1
Detailed description of the coding scheme

The coding scheme consists of the following six variables:

Variable:	Which indicates:
ACTOR	the producer of an utterance
EXCHANGE	the type of information that is communicated in a general sense
DISTANCE	relevance of the utterance to the question from the questionnaire
SPECIFICATION	further specification of the type of information
ADEQUACY	evaluation with respect to standardization
DIRECTION	chosen response alternative(s)

Every utterance is coded for these variables, and for each variable several codes are possible. Table 2, at the end of this appendix, shows all possible values for the six code variables. Below we will discuss the codes for each variable.

ACTOR can take the following values:

- I: the *interviewer*
- R: the *respondent*
- P: a *third person*

EXCHANGE can take the following values:

- Q: *Questions,* which are intended to obtain information that is required by the questionnaire.
- A: *Answers,* which are provisions of information that is required by the questionnaire.
- P: *Perceptions,* which are utterances that indicate that the utterance of the other party has been perceived.
- R: *Requests,* which are intended to solve communication problems, either requesting the other party to repeat, or to clarify their last utterance, or a request of the interviewer to look at the show card.
- C: *Comments*, which are used to give some qualification of an utterance (e.g., "difficult question", or "that's nice").
- D: *Detours,* which concern utterances that are not related to the ongoing conversation (e.g., "Would you like a cup of tea?").
- u: *Unintelligible,* which means that the utterance is indecipherable due to tape interruptions, background noises etc

The variable DISTANCE indicates to what extent utterances are related to the question from the questionnaire. It can take the following values:

- 0: behavior that is directly related to the questions from the questionnaire.

- 1: elaborations and motivations of (answers to) questions from the questionnaire.
- 2: further specifications of those behaviors.
- 3: entirely irrelevant utterances.
- B: backward, i.e., behaviors that refer to previous Q-A sequences in the interview.
- F: forward, i.e., behaviors that refer to upcoming questions.
- G: general, i.e., not referring to Q-A sequences or questions but to general aspects of the survey.

In Example 1, a Q-A sequence is shown, illustrating code categories for the three variables described so far.

Example 1 Q-A sequence coded for ACTOR, EXCHANGE and DISTANCE.

	ACTOR	EXCHANGE	DISTANCE	Verbal utterances	
1	I	Q	0	I: Do you have a very good, good, reasonable or bad health?	I poses question as worded
2	R	R	0	R: Could you repeat that?	R asks I to repeat
3	I	Q	0	I: Do you have a very good, good, reasonable or bad health?	I poses question as worded
4	R	A	1	R: Well I do visit my G.P. very often	R gives indirect answer
5	I	P	1	I: Uh huh	I perceives indirect answer
6	R	A	0	R: But my health is reasonable	R gives direct answer
7	I	P	0	I: Reasonable	I repeats direct answer
8	R	A	2	R: I have a very good G.P. you know	R elaborates his indirect answer
9	I	P	0	I: So reasonable	I repeats direct answer

The variable SPECIFICATION may be used to further specify the category coded for EXCHANGE. The variable ADEQUACY evaluates the utterances with respect to standardization. Both variables will be described below within the categories of EXCHANGE

Questions

Questions receive the code 'IQ0' for the first three variables.

The specification of questions can take the following values:

- C: a closed or *choice* question has a prescribed list of response alternatives from which the respondent has to choose one or more alternatives, or there is a prescribed answering format that is implied with the question (i.e., with implicit

alternatives, for example "How many days a week do you generally watch television?").

- O: an *open* question has no response alternatives and no prescribed answering format.
- Y: a *yes-or-no* question is a closed question that can be answered only with yes, no, (or, if applicable, 'don't know' or 'refused').
- S: an *assertion* is a statement, for which respondents are required to state whether, and often to what extent, they agree or disagree. This code was not part of the original coding scheme (Dijkstra 1999), but was added to distinguish assertions from yes-no questions and choice questions.
- A: response *alternatives* is a list of prescribed response alternatives.
- I: *introduction* is a scripted introduction to a question, e.g., "Now I will ask some questions about…".
- D: *definition* is a scripted definition that belongs to a question. This code was not part of the original coding scheme (Dijkstra 1999), but was added to distinguish scripted definitions from introduction text.
- M: *meaning* of question is clarification of the question.
- 0: *not posed,* i.e the Q-A sequence is empty because the question was not posed.

The ADEQUACY of question reading can take the following values:

- A: adequate questions: read as worded in the questionnaire.
- I: invalid questions, i.e., changed with respect to its original meaning.
- M: mismatch questions, i.e., changed with respect to literal wording (for example because interviewers use synonyms) but not changed with respect to meaning.
- S: suggestive questions, i.e., suggesting one or a few of the response options.

Answers

The SPECIFICATION of answers can take the following values:

- A: choosing an alternative from the explicit or implicit list of alternatives, i.e., the format of the response fulfills the requirement of the closed question, but does not have to be precisely formatted.
- O: an open answer, this is an answer by means of which the respondent does not choose a response alternative.
- Y: choosing yes or no.
- b: a don't know answer.
- r: a refusal.

The ADEQUACY of answers can take the following values:

- A: adequate, the answer is directly scorable, and there is no misunderstanding possible what alternative is meant.

- I: invalid, i.e., indicating misunderstanding of the question.
- M: mismatch answers, i.e., answers that are not formatted according to the prescribed alternatives, it is not exactly clear what alternative is meant.
- T: qualified answers, i.e., answers that indicate uncertainty about the accuracy (like 'I think', or 'I am not sure but'). This code was not part of the original coding scheme (Dijkstra 1999), but was added to distinguish qualified answers from mismatch answers and adequate answers.

Perceptions

The SPECIFICATION of perceptions can take the following values

- E: echoes other party, by means of complete or partial repeat or paraphrase.
- n: notes other party, with 'uhuh' or 'yes'.
- f: filled pause, 'uh's' or meaningless utterances like 'well'.
- s: silence, i.e., a meaningful silence, in between the utterances of the same speaker. For example, after the reading of the question, the respondent can remain silent very long, which triggers the interviewer to repeat the question.

The ADEQUACY of perceptions can take the following values, when the category 'Echo' is coded for perception:

- A: adequate,
- M: mismatch, i.e., the repetition is not a literal echo of the utterance but a paraphrase.

Requests, comments and detours

The SPECIFICATION of *requests* can take the following values

-d: requests to the other party to repeat.

-m: requests to the other party to clarify the meaning of the utterance.

-o: other requests, e.g., "Shall I repeat that?" or "Can you speak a little bit slower?"

-A: requests to the other party to look at the show card with response alternatives.

The SPECIFICATION of *comments* can take the following values

-t: task oriented, e.g., comments from the interviewer like "you are doing okay" or comments from the respondent such as "what a silly question".

-p: personal oriented, e.g., comments from the interviewer like "I have a bad health too"

The SPECIFICATION of *detours* can take the following values

-t: task oriented, e.g., utterances like "What will happen with my answers?"

-p: personal oriented, e.g., utterances like "Would you like a cup of tea?" or "How long have you been an interviewer?"

With the variable DIRECTION it is possible to code for the *number* of the particular alternative that was mentioned (i.e., was it the first, second, etc., of the prescribed list of alternatives). This information can be useful, for example to investigate whether respondents accept the alternatives offered by interviewers. The code for direction can also be compared with the score as entered by the interviewer.

When utterances do not refer to a single alternative the following values are possible:
- a: all alternatives (a)
- l: low, the first alternatives of a set
- m: middle, the middle alternatives of a set
- h: high, the last alternatives of a set
- s: subset, an unordered subset of alternatives
- z: no alternatives
- x: it is not relevant to mention alternatives
- I: interrupted

The meaning of DIRECTION for questions

By means of DIRECTION it is possible to ascertain whether alternatives were read *within, after or before* the question. For example, a closed choice question with the alternatives incorporated (e.g., 'Do you own one, two, more than two or no bicycles at all?') receives the code 'IQ0CAa' ('Interviewer reads closed question with all alternatives'). An interviewer may reword such a question by reading the alternatives after the question proper (e.g., 'How many bicycles do you own, is that one, two, more than two or none?'). This question reading is counted as two separate utterances that receive the codes 'IQ0CAz' ('Interviewer reads closed question without alternatives') and 'IQ0AAa' ('Interviewer reads all alternatives') respectively. This distinction enables studies on the effects of the place of the question component on for example the occurrence of interruptions.

The meaning of DIRECTION for answers

By means of DIRECTION it is possible to code for response precision, which may indicate the seriousness of mismatch answers. This distinction entails answers that unequivocally point at a response alternative, but are not formatted as such, answers that seem to point at a subset of the response alternatives, and answers that seem to point at no particular response alternative. For example, in case of a five-point scale with the response alternatives (1) Strongly agree, (2) Agree, (3) Neither agree nor disagree, (4) Disagree and (5) Strongly disagree, respondents can give multiple types of answers, as is summarized in Table 1. In example 2, a Q-A is shown that is coded for all variables.

Table 1 Examples of adequate and mismatch answers for a 5-piont agree disagree scale

Utterance	*Points at*	*ADEQUACY*	*DIRECTION*
R: I do agree with that	'agree'	adequate	2 (2nd alternative)
R: In between	'neither agree nor disagree'	mismatch	3 (3rd alternative)
R: I certainly agree	'strongly agree' or 'agree'	mismatch	l (low alternatives)
R: I do not disagree	'agree' or 'neither agree nor disagree'	mismatch	m (middle alternatives)
R: Sometimes	None of the alternatives	mismatch	x (no alternative)

Example 2 Q-A sequence coded for all variables

	ACTOR	EXCHANGE	DISTANCE	SPECIFICATION	ADEQUACY	DIRECTION	Verbal utterances	Explanation
1	I	Q	0	C	A	a	I: Do you have a very good, good, reasonable or bad health?	I poses question as worded adequately with all alternatives
2	R	R	0	d	x	x	R: Could you repeat that?	R asks I to repeat
3	I	Q	0	C	S	h	I: Do you have a good, reasonable or bad health?	I poses question suggestively: only alternatives 2, 3 and 4.
4	R	A	0	A	M	h	R: Not so bad	R gives mismatch answer that refers to options 2, 3 and 4
5	I	Q	0	A	A	h	I: Would you say good, reasonable or bad?	I repeats alternatives 2, 3 and 4.
6	R	A	0	A	A	3	R: Well, reasonable	R gives direct answer, that is adequate and refers to the 3rd response option
7	I	P	0	E	A	3	I: Reasonable	I adequately repeats direct answer referring to option 3
8	R	A	1	O	x	x	R: Compared to my wife yes	R elaborates his answer
9	I	P	0	E	A	3	I: So reasonable	I repeats direct answer

Table 2 Overview of all variables and all categories of the scheme.

Actor	**Exchange**	**Distance**	**Specification**	**Adequacy**	**Direction**
Interviewer Respondent Third Person	Q: question	0: from script 1: related to 0 2: related to 1 3: irrelevant G: general F: forward B: backward	O: open question C: closed Q Y: yes/no Q S: statement A: alternatives I: introduction D: definition M: meaning of Q 0 : not posed	A: adequate M: mismatch I: invalid S: suggestive	a: all i: interruption l: low m: middle h: high s: subset 0,1,2…9 z: no direction
Interviewer Respondent Third Person	A: answer	0: from script 1: related to 0 2: related to 1 3: irrelevant	Y: yes/no A A: alternative O: open answer	A: adequate M: mismatch I: invalid T: qualified	l: low m: middle h: high 0,1,2...9
		G: general F: forward	b: don't know r: refusal	x x	x x
		B: backward			
Interviewer Respondent Third Person	P: perception	0: from script 1: related to 0 2: related to 1 3: irrelevant G: general	E: echo (repeats other)	A: adequate M: mismatch	a: all l: low m: middle h: high 0,1,2...
		F: forward B: backward	n: notes other f: filled pause s: silence	x x x	x x x
Interviewer Respondent Third Person	R: request	0,1,2,3 G, B, F	d: duplicate (request repetition) m: meaning o: other A: show card	x x x	x x x
Interviewer Respondent Third Person	C: comment	0,1,2,3 G, B, F	p: personal t: task	x x	x x
Interviewer Respondent Third Person	D: detour	G: general	p: personal t: task	x x	x x
Interviewer Respondent Third Person	u: unintelligeble	u	u	u	u

Appendix 5-1
38 questions of the CATI survey interview

1. I would like to ask you a few questions about watching television first. How many days per week do you watch television, on average?
7. each day, 6. of seven days, 5. of seven days,4. of seven days
3. of seven days,2. of seven days,1. of seven days,0. never (to Q 20)
(FOR **ALL** QUESTIONS TWO ANSWER POSSIBILITIES FOR 'REFUSAL' OR 'DON' T KNOW' WERE ALSO AVAILABLE)

2. And when you do watch television, for how many hours or minutes are you doing this, on average
... hours + ... minutes

3. When was the last time you were watching television?
1. Monday,2. Tuesday,3. Wednesday,4. Thursday,5. Friday,6. Saturday
7. Sunday,8. a week or longer ago

5. Did you at that time switch on the television set out of interest for a programme or out of pastime?
3. both, 2. more out of interest, 1. more out of pastime, 0. neither interest nor pastime

6. Where did you watch television?
1. at home, 2. at work, 3. with friends/family/ acquaintance and the like, 4. waiting room G.P., hospital and the like, 5. cafe/restaurant
6. transportation (train/bus/aeroplane),7. other

7. Were you watching television with others?
1. yes with others, 0 no, alone

8. For how many hours or minutes were you watching? Please give an estimation!
... hours + ... minutes

9. What percentage of the time were you watching attentively?
... percent
Allowed explanation for Q9:
100% is all of the time, 0% is none of the time, 50% is half of the time

10. You were just describing the number of hours/minutes you were watching television. How many blocks of commercials did you see during this period? Please give an estimation.
... blocks, 0: to Q 13

11. How many commercials did you see? Please give an estimation.
... commercials, 0: did not see any commercials (to Q13)

12. What percentage of these commercials did draw your attention?
... percent

13. You were describing the specific moment you were watching television before. The next questions however concern television advertising in general.
I will mention some possible reactions to television advertising. Would you please tell me whether you do this each time, often, sometimes or never when commercials appear on the screen?
You stay to watch the commercials
4. each time,3. often,2. sometimes,1. never

14. You zap to another channel.
4. each time,3. often,2. sometimes,1. never

15. You switch off the volume.
4. each time,3. often,2. sometimes,1. never

16. You switch off the television set.
4. each time,3. often,2. sometimes,1. never

17. You do something else but leave the commercials on.
4. each time,3. often,2. sometimes,1. never

18. You leave the room.
4. each time,3. often,2. sometimes,1. never

19. You even search for the commercials.
4. each time,3. often,2. sometimes,1. never

20. Now I'll read some statements about television advertising. We would like to know your opinion about these statements. Would you please indicate whether you agree or disagree with the statement.
I think commercials on television provide me with useful information about special offers.
1. strongly disagree, 2. disagree, 3. neither agree nor disagree
4. agree, 5. strongly agree

[Instruction for Q 20-29: After the first answer 'agree' or 'disagree' ask whether the respondent strongly (dis)agrees or just (dis)agrees.]

21. For me commercials are funny.
1. strongly disagree, 2. disagree, 3. neither agree nor disagree
4. agree, 5. strongly agree

22. For me, commercials on television provide me with meaningful information about the product use of consumers.
1. strongly disagree, 2. disagree, 3. neither agree nor disagree
4. agree, 5. strongly agree

23. For me, commercials provide me with useful information about new products.
1. strongly disagree, 2. disagree, 3. neither agree nor disagree
4. agree, 5. strongly agree

24. For me, television commercials are entertaining
1. strongly disagree, 2. disagree, 3. neither agree nor disagree
4. agree, 5. strongly agree

25. For me, television commercials appear at inconvenient moments
1. strongly disagree, 2. disagree, 3. neither agree nor disagree
4. agree, 5. strongly agree

26. For me, television commercials are too blaring.
1. strongly disagree, 2. disagree, 3. neither agree nor disagree
4. agree, 5. strongly agree

27. For me, television commercials are implausible.
1. strongly disagree, 2. disagree, 3. neither agree nor disagree
4. agree, 5. strongly agree

28. For me, television commercials are repeated too often.
1. strongly disagree, 2. disagree, 3. neither agree nor disagree
4. agree, 5. strongly agree

29. For me, television commercials are too much alike.
1. strongly disagree, 2. disagree, 3. neither agree nor disagree
4. agree, 5. strongly agree

30. When you are watching television, do you always, often, sometimes or never pay attention to the commercials?
4. each time, 3. often,2. sometimes,1. never

31. During the next part of the questionnaire I would like to ask you some questions about advertising in different media, such as television, radio, newspapers and magazines.
Are you in general negative or positive towards television advertising?
1. strongly negative, 2. negative, 3. neutral,4. positive,5. strongly positive

[Instruction for Q 31-35: After the first answer 'positive' or 'negative' ask whether this is very positive/negative or just positive/negative.]

32. Are you in general negative or positive towards radio advertising?
1. strongly negative, 2. negative, 3. neutral,4. positive,5. strongly positive

33. Are you in general negative or positive towards newspaper advertising?
1. strongly negative, 2. negative, 3. neutral,4. positive,5. strongly positive

34. Are you in general negative or positive towards magazine advertising?
1. strongly negative, 2. negative, 3. neutral,4. positive,5. strongly positive

35. Please give an overall judgement: are you in general negative or positive towards advertising?
1. strongly negative, 2. negative, 3. neutral,4. positive,5. strongly positive

44. Finally, I would like to ask some general questions. How many persons does the household that you are part of include? What matters is the household at this moment, so please do not count children that do not live at home. You should include yourself in the count.
... persons

45. What is your age?
... years old

46. Are you working at this moment, that is employed, in a company, with the government or self-employed?
1. yes, 0. no (to Q 48)

49. What is the highest type of education you have enjoyed, either or not completed with a cartificate?
1. Elementary education (Dutch: 'basisschool'),2. Technical and vocational training for 12-16 year-olds (Dutch: 'lager beroepsonderwijs':LTS, LEAO, LHNO, huishoudschool, VBO), 3. Secondary education (not pre-university)(Dutch: 'middelbaar algemeen'), 4. Technical and vocational training for 16-18 year-olds (Dutch: 'middelbaar beroeps', 5. Secondary education (pre-university) (Dutch: 'voortgezet algemeen': HAVO, VWO, MMS, HBS, etc.), 6. Technical and vocational training for 18+ (Dutch 'hoger beroepsonderwijs',HTS, HEAO, div. academies, MO, etc.), 7. University (Dutch: 'universiteit')
8. Other

Appendix 6-1

Recognition of three types of mismatch answers

The mismatch answers were divided into the three different kinds of mismatch answers: conversational, task and cognitive mismatch answers. Below we will give examples of how the three types of mismatch answers were recognized.

Conversational mismatch answers

Conversational mismatch answers occur when the information required by the question is readily available, and the respondent is not faced with a problematic or ambiguous response task. The respondent does not force himself to get engaged into time-consuming processes to retrieve the response alternatives that were read by the interviewer. These mismatch answers are very likely to occur immediately after the interviewer posed the question, i.e., without hesitations, preceding requests for clarification, or verbal considerations.

It must be noted that many conversational mismatch answers are not highly problematic, because respondents often spontaneously give an adequate answer immediately after their mismatch answer, not requiring any probing or clarification from the interviewer.

Task mismatch answers

Task mismatch answers occur when the respondent is faced with some ambiguity in the response task. Task mismatch answers are often preceded by (indirect) requests for clarification.

An example of a task mismatch answer is shown in Excerpt 1. This question concerns the respondent's association about a political party. The respondent has some difficulty in selecting a response alternative. He is not satisfied about the third alternative 'hardly associated', and gives the mismatch answer 'very little'. He ends up with the mismatch 'a little bit more', meaning something more than 'not at all', but not clearly indicating 'hardly'. Such a task mismatch gives information about possible improvements of the listed response alternatives.

Excerpt 1 Q-A sequence concerning ESS question C30

1.	I:	To what extent do you consider yourself associated with this party?
2.	I:	Do you feel very associated, fairly associated, hardly associated or not at all associated?
3.	R:	Well, hardly, very little at least
4.	I:	fairly or hardly?
5.	I:	hardly?
6.	R:	what's the least?
7.	I:	not at all associated
8.	R:	no a little bit more
9.	I:	so three

Cognitive mismatch answers

Cognitive mismatch answers may occur when the respondent is faced with a difficult response task that requires a relatively high amount of cognitive processing. Verbal comments that the information required to answer the question is not readily available may indicate this. While respondents are processing the information, they may give verbal considerations or verbal enumerations, as is shown in Excerpt 2. The verbal expression in line 2 is an inference the respondent uses for his estimation. This is a strategy respondents may use to avoid retrieving complex information. In this case, the respondent does not express problems with the meaning of the question, but with the accuracy of the estimation. The answer he gives in line 3 is a mismatch answer because two answers are given, and it is not clear which of the answers is the eventual response. The mismatch answer may even be part of the estimation strategy the respondent is carrying out (i.e., the respondent is not finished calculating). Interestingly, the interviewer immediately takes the first part of the respondent's mismatch answer ('two hours'), and does not stimulate the respondent to retrieve more information to solve the mismatch answer adequately.

Excerpt 2 Q-A sequence concerning ESS question B2

1.	I:	Uh on an average day, how long do you watch political news and current affairs?
2.	R:	That is I think it is about half of what I watch
3.	R:	So that must be two hours or so or two and a half
4.	I:	Two hours

Appendix 7-1
Wording of all versions of the questions in Dutch and English

GUIDELINE TO THIS APPENDIX

The questionnaire is divided into nine topic sections: I Perceived health; II GHPQ assertions (General Health Perception Questionnaire); III Spare time; IV Government and health assertions; V Food habits; VI Public health assertions; VII Health contacts; VIII Body measures, IX Background questions.

Questions were manipulated for one or multiple hypotheses, and were literally derived or adapted from original wording of actual surveys, or question wording was entirely new.

	Literal wording	Adapted wording	New question
H1: Conversational versus formal questions	2, 3, 4, 5, 7, 18, 19, 27, 28, 29, 30, 31, 35 (both versions), 39 (both versions), 40, 41 (both versions)	1, 6, 8, 9, 10, 16, 20, 32, 36, 42, 43	11, 12, 13, 14, 17, 21, 22, 23, 24, 26, 33, 34, 37, 38
H2a: Conversational versus formal alternatives	2, 3, 4, 5, 7, 18, 19, 27, 28, 29, 30, 31	6, 8, 9, 16, 20, 32	15, 21, 37, 38
H2b: Implicit versus listed alternatives		16	21, 24, 33, 34
H3: Difficult versus easy questions			12, 14
H4: Ambiguous versus non-ambiguous questions	40	1, 10, 43	12, 17, 27

Below, the introduction statements and exact question wordings of all questions within the nine sections are presented. The information of the questionnaire is arranged within three columns:

- In the first column the question number is given. These numbers are used in chapter 7.
- In the second column, the type of manipulation is given. The abbreviations refer to the four manipulations: conversation-likeness of response alternatives and question wording, difficulty and ambiguity (see table 7-1 in chapter 7). It is indicated for all question wordings whether they concern formal or conversational question wording and whether the response alternatives for the questions are formal, conversational or implicit. The manipulation of difficulty and ambiguity is only indicated when applicable (i.e., when difficult or ambiguous versions were included for the specific question). Therefore, when no information in column 2 is provided with respect to difficulty and ambiguity of the questions, it concerns easy and non-ambiguous questions only.
- In the third column the question wording of each version of a question is provided in Dutch and in English (in italics).

INTRODUCTION TEXT OF THE INTERVIEW		
Goedenavond, u spreekt met … van de Vrije Universiteit Amsterdam. Wij zijn bezig met een onderzoek over voeding, vrijetijdsbesteding en gezondheid. Ik zou u graag hierover een paar vragen willen stellen. *Good evening, this is... from the Free University at Amsterdam. We are conducting a survey on food, spare time use and health. I would like to ask you some questions about this.*		
STATEMENT AFTER FIRST APROVAL OF COOPERATION		
Fijn dat u mee wilt werken. Zoals ik zei gaat het interview over voeding, vrijetijdsbesteding en gezondheid. Het interview zal hooguit tien minuutjes duren en ik wil dan nu graag beginnen met de eerste vraag. *How nice you will cooperate. As I mentioned the interview will concern food, spare time use and health. The interview will last no more than ten minutes and I you agree I would like to start with the first question.*		
SECTION I Perceived health		
Qnr	type of manipulation	Question wording
1		**Introduction statement (Questionnaires 1 to 4):** Ik wil graag beginnen met enkele vragen over gezondheid. *I would like to start with some question about health.*
1	CFEN **-Formal Q** -Conv. Alt, **-Non-ambiguous**	Beschikt u, vergeleken met uw leeftijdsgenoten, over een zeer goede, goede, redelijke of slechte lichamelijke gezondheidstoestand? *As compared to your peers, would you consider your physical health as very good, good, reasonable or bad?*
1	CCEN **-Conv. Q** -Conv. Alt **-Non-ambiguous**	Heeft u voor uw leeftijd een zeer goede, goede, redelijke of slechte gezondheid? *According to your age, do you have a very good, good, reasonable or bad health?*
1	CFEA **-Formal Q** -Conv. Alt **-Ambiguous**	Beschikt u over zeer goede, goede, redelijke of slechte gezondheidstoestand? *Would you consider your health as very good, good, reasonable or bad?*
1	CCEA **-Conv. Q** -Conv. alt. **-Ambiguous**	Heeft u een zeer goede, goede, redelijke of slechte gezondheid?[15] *Do you have a very good, good, reasonable or bad health?*

[15] Adapted from original Dutch wording "Hoe is over het algemeen uw gezondheidstoestand? zeer goed, goed, gaat wel, soms goed, soms slecht, of slecht?" (CBS Health Survey, see Snijkers, 1999).

SECTION II GHPQ-assertions		
2	Conversational alternatives	**Introduction statement:** Ik ga nu een aantal stellingen aan u voorleggen. Om aan te geven of iets voor u wel of niet van toepassing is kunt u hierop antwoorden met 'JA' , 'MISSCHIEN' of 'NEE' Now I am going to read some assertions. To indicate whether something applies for you, you can answer with 'YES' MAYBE' or 'NO'
2	Formal alternatives	**Introduction statement:** Ik ga nu een aantal stellingen aan u voorleggen. Om aan te geven of iets voor u wel of niet van toepassing is kunt u hierop antwoorden met 'WAAR' , 'MOGELIJK WAAR' of 'NIET WAAR' *Now I am going to read some assertions. To indicate whether something applies for you, you can answer with 'TRUE' 'POSSIBLY TRUE' or "FALSE'*
2	(All versions)	**Second introduction statement:** De eerste stelling luidt: *The first assertion is as follows:*
2	Formal assertion	Mijn gezondheid baart mij zorgen. *My health causes me worries*
2	Conversational assertion	Ik maak me zorgen over mijn gezondheid[16]. *I worry about my health*
3	(All versions)	**Introduction statement:** De volgende stelling is: *The next assertion is:*
3	Formal assertion	Sporadisch ziek worden behoort bij het leven. *Getting sick once in a while is part of life*
3	Conversational assertion	Af en toe ziek worden hoort bij het leven.[17] *Getting sick sometimes is part of life*
4	Formal assertion	Ik heb de afgelopen periode een slechte lichamelijke gesteldheid gehad. *Recently I have had a poor physical health*
4	Conversational assertion	Ik voel me de laatste tijd slecht.[18] *I have been feeling bad lately*
5	Formal assertion	Mijn gezondheidstoestand is uitstekend[19]. *My health is excellent*
5	Conversational assertion	Ik ben zo gezond als een vis. *I am as healthy as an ox*

[16] Adapted from original (conversational) Dutch wording "Ik maak me nooit zorgen over mijn gezondheid" (Kriegsman, Van Eijk and Deeg, 1995). The word 'nooit' ('never') was deleted from this original question wording, because it is obviously problematic to use negations in question wording.
[17] Taken from original (conversational) Dutch wording (Kriegsman, Van Eijk and Deeg, 1995)
[18] Taken from original (conversational) Dutch wording (Kempen et al. 1995)
[19] Adapted from original (formal) Dutch wording "Mijn gezondheid is uitstekend" (Kempen et al. 1995).

6	Formal assertion	Ik aanvaard dat ik van tijd tot tijd nu eenmaal ziek word.[20] *I acknowledge that from time to time I will be sick*
6	Conversational assertion	Ik accepteer het dat ik soms gewoon ziek word *I believe that sometimes I am just going to be sick*
7	Formal assertion	Het komt me voor dat ik gemakkelijker ziek word dan andere mensen *It appears to me that I get sick easier than other people*
7	Conversational assertion	Ik lijk wat makkelijker ziek te worden dan andere mensen[21] *I seem to get sick a little easier than other people*
8	Conversational assertion	Ik ben ziek[22] *I am ill*
9	Formal assertion	Ik ben een enkele maal zo ziek geweest dat ik dacht te overlijden *I have on one occasion been so sick I thought I would pass away*
9	Conversational assertion	Ik ben wel eens zó ziek geweest dat ik dacht dat ik doodging[23] *I was so sick once I thought I might die*
SECTION III Spare time Choice questions		
10		**Introduction statement (questionnaires 1 to 4):** Nu volgen enkele vragen over vrijetijdsbesteding. *Now some questions about your spare time use.*
10		**Introduction statement (questionnaires 5 and 6):** De eerste vragen gaan over vrijetijdsbesteding. *The first questions will be about your spare time use.*
10		**Definition preceding non-ambiguous, formal question:** Met sportieve activiteiten bedoelen we lichamelijk inspannende sporten als vrijetijdsbesteding. *By sporty activities we mean physically intenive sports ase leisure activity*
10	IFEN/IFEA **-Formal Q** -Implicit alternatives **-Ambiguous/ non-ambiguous**	Heeft u de afgelopen 12 maanden aan sportieve activiteiten gedaan?[24] *Have you during the past 12 months been engaged in any sporty activities?*

[20] Adapted from original (formal) Dutch wording "Ik aanvaard dat ik soms nu eenmaal ziek word" (Kriegsman, Van Eijk and Deeg, 1995)
[21] Taken from original (conversational) Dutch wording (Kriegsman, Van Eijk and Deeg, 1995)
[22] Adapted from original (conversational) Dutch wording "Ik ben een beetje ziek" (Kempen et al. 1995).
[23] Adapted from original (conversational) Dutch wording "Ik ben wel eens zó ziek geweest dat ik dacht dat ik wel dood kon gaan" (Kriegsman, Van Eijk and Deeg, 1995)
[24] Adapted from original (formal) question wording ("Doet u op dit moment aan sportieve activiteiten of heeft u dat het laatste jaar gedaan?" CAPI questionnaire Familie-enquête beroepsbevolking 1998: http://www.niwi.knaw.nl/nl/maatschappijwetenschappen/steinmetzarchief/dddi/docs/p1583.pdf/p1583.pdf).

		Definition preceding non-ambiguous, conversational question: Met sport bedoelen we lichamelijk inspannende sporten in uw vrije tijd. *By sport we mean physically intensive sports in your spare time*
10	ICEN/ICEA **-Conv. Q** -Implicit alternatives **-Ambiguous/ non-ambiguous**	Heeft u de laatste 12 maanden aan sport gedaan? *Have you been engaged in any sport during the last 12 months?*
11	(all versions)	**Introduction statement:** Bij de volgende vragen is het begrip week belangrijk Wanneer we spreken over een week bedoelen we alle 7 dagen van de week, dus zowel het weekend als doordeweeks. *For the next question the concept of 'week' is important. When we speak of a week we mean all 7 days of the week, so both weekdays and weekend.*
11	IFEN/IFEA **-Formal Q** -Implicit alternatives -Non-ambiguous	Op hoeveel dagen tijdens de afgelopen week deed u aan sport? *On how many days during the past week have you been engaged in any sports*
11	ICEN/ICEA **-Conv. Q** -Implicit alternatives -Non-ambiguous	Op hoeveel dagen heeft u vorige week gesport? *How many days did you sport last week?*
12	IFEN/IFEA/ IFDN/IFDA **-Formal Q** -Implicit alternatives -**Difficult/ Easy** **-Ambiguous/ non-ambiguous**	Wat is het totale aantal uren en minuten dat u in die week aan sport besteed heeft? *What is the total number of hours and minutes that you spent on sports in that week?*
12	ICEN/ICEA/ ICDN/ICDA **-Conv. Q** -Implicit alt. -**Difficult/ Easy** **-Ambiguous/ non-ambiguous**	Hoeveel uren en minuten was u vorige week in totaal kwijt aan sporten? *How many hours and minutes did you spend on sports?*

13	IFEN **-Formal Q** -Implicit alt.	Wandelt of loopt u wel eens minstens 10 minuten aaneen? *Do you ever hike or walk for ten minutes consecutively?*
13	ICEN **-Conv. Q** -Implicit alt.	Wandelt of loopt u wel eens 10 minuten of meer achterelkaar? *Do you ever hike or walk 10 minutes or more on end?*
14	IFEN **-Formal Q** -Implicit alt. **-Easy**	Wat is het aantal dagen dat u dat afgelopen week heeft gedaan? *What is the number of days that you did this in the past week?*
14	ICEN **-Conv. Q** -Implicit alt. **-Easy**	Op hoeveel dagen heeft u dat afgelopen week gedaan? *How many days did you do this in the past week?*
14	IFDN -**Formal Q** -Implicit alt. **-Difficult**	Wat is het aantal dagen dat u de afgelopen week minstens 10 minuten aaneen heeft gewandeld of gelopen? In the past week, what is the number of days that you did walk for at least 10 minutes consecutively hike or walk?
14	ICDN -**Conv. Q** -Implicit alt. **-Difficult**	Op hoeveel dagen heeft u vorige week 10 minuten of meer achterelkaar gewandeld of gelopen? How many days did you walk or hike for ten minutes or more on end?
15	CFEN -Formal Q **-Conv. alt.**	Wandelt of loopt u dan meestal in een rustig, gewoon of snel tempo? *Do you usually walk at a slow, regular or fast pace?*
15	FFEN -Formal Q **-Formal alt.**	Wandelt of loopt u dan meestal in een laag, middelmatig of hoog tempo? *Do you usually walk at a low, mediate or high pace?*
16	IFEN **-Formal Q** **-Implicit alt.**	Wat is het aantal dagen per week, dat u doorgaans televisie kijkt? *What is the number of days a week that you usually watch television?*
16	ICEN **-Conv. Q** **-Implicit alt.**	Hoeveel dagen per week kijkt u meestal TV? *How many days a week do you usually watch TV*?
16	CCEN **-Conv. Q** **-Conv. alt.**	Kijkt u elke dag, de meeste dagen, sommige dagen, of bijna nooit TV? *Do you watch television every day, most days, some days or hardly ever?*
16	FCEN **-Conv. Q** **-Formal alt.**	Kijkt u praktisch 0 dagen, 1 tot 3 dagen, 4 tot 6 dagen of 7 dagen per week TV? *Do you practically watch 0 days, 1 to 3 days, 4 to 6 days or 7 days a week TV?*

17		**Definition preceding non-ambiguous, formal question:** Bij de volgende vraag gaat het om tv kijken met volle aandacht. *The next question concerns watching television with full attention*
17	IFDN **-Formal Q** -Implicit alt. -Difficult **-Ambiguous/ non-ambiguous**	En, op de dagen dat u kijkt, wat is op 1 dag het totale aantal uren of minuten dat u doorgaans televisie kijkt? *And, on the days that you watch, what is on one day the total number of hours and minutes that you usually watch television?*
17		**Definition preceding non-ambiguous, formal question:** En nu gaat het om tv kijken met volle aandacht. *Now only watching television with full attention is considered*
17	ICDN **-Conv. Q** -Implicit alt. -Difficult **-Ambiguous/ non-ambiguous**	En op de dagen dat u kijkt, hoeveel uren of minuten bij elkaar kijkt u dan meestal op 1 dag? *And, on the days that you watch, how many hours and minutes in all do you usually watch television?*
SECTION IV Government and health assertions		
18	Conversational alternatives	**Introduction statement:** Ik ga nu een aantal stellingen aan u voorleggen. Om aan te geven of u het wel of niet eens bent met de stelling kunt u hierop antwoorden met 'JA' of 'NEE' *Now I am going to read some assertions. To indicate whether or not you agree with the assertion you can answer with 'YES' or 'NO'*
18	Formal alternatives	**Introduction statement:** Ik ga nu een aantal stellingen aan u voorleggen. Om aan te geven of u het wel of niet eens bent met de stelling kunt u hierop antwoorden met de volgende vijf antwoordmogelijkheden 'ZEER MEE EENS', 'MEE EENS', 'NEUTRAAL', 'ONEENS', 'ZEER MEE ONEENS'. *Now I am going to read some assertions. To indicate whether or not you agree with the assertion you can answer with the following five answer possibilities 'STRONGLY AGREE', 'AGREE' 'NEUTRAL', 'DISAGREE', 'STRONGLY DISAGRRE'*
18	Conversational assertion	De regering moet er voor zorgen dat er voor de mensen rookvrije cafés en restaurants zijn. *The government should take care of smoke free cafes and restaurants for the people*
18	Formal assertion	De overheid dient er voor zorgen dat burgers gebruik kunnen maken van rookvrije horeca[25] *The government should provide civilians with a smoke free hotel and catering industry*

[25] Taken from original (formal) Dutch wording (RVD 2003)

19	Conversational assertion	Ik vind dat de overheid voorlichting aan de mensen moet geven over de gevolgen van roken en meeroken *I think the government should educate the people about the consequences of smoking and passive smoking*
19	Formal assertion	Het is nodig dat de overheid de bevolking voorlicht over de gevolgen van roken en meeroken[26] *It is necessary that the government educates civilians about the consequences of smoking and passive smoking*
20	Conversational assertion	Volgens mij wordt ons eten goed gecontroleerd zodat wat in de winkel ligt veilig is. *I think our food is well checked so what is in stores is safe*
20	Formal assertion	Ik vertrouw erop dat er goed wordt gecontroleerd, zodat winkels zijn voorzien van veilig voedsel.[27] *I trust that inspections are good so stores are supplied with safe food*
SECTION V Food habits choice questions		
21	(All versions)	**Introduction statement:** Nu volgen enkele vragen over uw voedingsgewoonten We hebben het daarbij steeds over de 5 doordeweekse dagen, dus zonder de dagen in het weekend. *Now I will ask you some questions about your food habits. We will talk only about the 5 weekdays, so without the days of the weekend*
21	IFEN **-Formal Q.** **-Implicit alt.**	Wat is het aantal doordeweekse dagen dat u graanproducten zoals brood, muesli of cornflakes bij het ontbijt gebruikt? *What is the number of weekdays that you use corn products such as bread, muesli or cornflakes as breakfast?*
21	ICEN **-Conv. Q.** **-Implicit alt.**	Hoeveel doordeweekse dagen heeft u graanproducten zoals brood, muesli of cornflakes als ontbijt? *How many weekdays do you use corn products such as bread, muesli or cornflakes as breakfast?*
21	FFEN **-Formal Q.** **-Formal alt**	Gebruikt u doordeweeks nooit, 1 tot 2 dagen per week, 3 tot 4 dagen per week of alle 5 dagen graanproducten zoals brood, muesli of cornflakes bij het ontbijt? *Do you on weekdays never, 1 to 2 days, 3 to 4 days or all 5 days use corn products such as bread, muesli or cornflakes as breakfast?*
21	CFEN **-Formal Q.** **-Conv. alt.**	Gebruikt u doordeweeks nooit, af en toe, de meeste dagen, of elke dag graanproducten zoals brood, muesli of cornflakes bij het ontbijt ? *Do you on weekdays never, once in a while, most days or every day use corn products such as bread, muesli or cornflakes as breakfast?*

[26] Taken from original (formal) Dutch wording (RVD 2003)
[27] Adapted from original (formal) Dutch wording "Ik vertrouw erop dat ons voedsel goed wordt gecontroleerd, zodat wat in de winkel ligt veilig is" (RVD 2003)

22	IFEN -**Formal Q.** -Implicit alt.	Wat is het aantal doordeweekse dagen dat u bij een warme maaltijd vlees gebruikt? *What is the number of weekdays that you use meat at dinner?*
22	ICEN **-Conv. Q.** -Implicit alt.	Hoeveel doordeweekse dagen eet u bij een warme maaltijd vlees? *How many weekdays do you use meat at dinner?*
23	(All versions)	**Definition statement preceding both versions of Q23** Bij de volgende vraag over vers fruit, bedoelen we alleen los fruit zoals appels of mandarijntjes en geen verse vruchtensappen. *In the next question about fresh fruits, we only include fruits such as apples and mandarins and no fresh fruit juices.*
23	ICEN **-Conv. Q.** -Implicit alt.	Op hoeveel dagen eet u doordeweeks meestal fruit? *On how many weekdays do you eat fruit?*
23	IFEN **-Formal Q.** -Implicit alt.	Wat is het aantal doordeweekse dagen dat u meestal fruit gebruikt? *What is the number of weekdays that you usually use fruit?*
24	IFDN **-Formal Q.** **-Implicit alt.** -Difficult	Nogmaals voor de doordeweekse dagen. Wat is het totale aantal koppen water, koffie, thee en andere non-alcoholische dranken dat u gewoonlijk per dag gebruikt? *Again for weekdays only. What is the total number of cups of coffee, tea and other non-alcoholic beverages that you usually use on a day?*
24	ICDN **-Conv. Q.** **-Implicit alt.** -Difficult	En weer voor de doordeweekse dagen. Hoeveel koppen water, koffie, thee en andere non-alcoholische dranken drinkt u meestal bij elkaar per dag? *Again for weekdays only. How many cups of coffee, tea and other non-alcoholic beverages do you usually drink in all on a day?*
24	FFDN **-Formal Q.** **-Formal alt.** -Difficult	Nogmaals voor de doordeweekse dagen. Gebuikt u per dag in totaal meer dan 8 koppen, ongeveer 8 koppen of minder dan 8 koppen water, koffie, thee en andere non-alcoholische dranken? *Again for weekdays only. During a day, do you use more than 8 cups, about 8 cups or less than 8 cups cups of coffee, tea and other non-alcoholic beverages?*
25	Filter question (all versions)	Gebruikt u doordeweeks en in het weekend wel eens een alcoholische drank? *On weekdays and during the weekend, do you use alcoholic beverages?*
26	ICDN -**Conv. Q** -Implicit alt. -Difficult	Hoeveel glazen alcohol drinkt u gemiddeld per week? *How many glasses of alcoholic beverages do you drink on average during a week?*
26	IFDN -**Formal Q**. -Implicit alt. -Difficult	Wat is het aantal glazen alcoholische drank dat u gemiddeld per week gebruikt? *What is the number of glasses of alcoholic beverages you drink on average during a week?*

SECTION VI Public health assertions		
27	Conversational alternatives	**Introduction statement:** Ik ga nu een aantal stellingen aan u voorleggen. Om aan te geven of u het wel of niet eens bent met de stelling kunt u hierop antwoorden met 'JA' of 'NEE' *Now I am going to read some assertions. To indicate whether or not you agree with the assertion you can answer with 'YES' or 'NO'*
27	Formal alternatives	**Introduction statement:** Ik ga nu een aantal stellingen aan u voorleggen. Om aan te geven of u het wel of niet eens bent met de stelling kunt u hierop antwoorden met 'MEE EENS', of 'ONEENS'. *Now I am going to read some assertions. To indicate whether or not you agree with the assertion you can answer with 'AGREE' or 'DISAGREE'*
27	Conversational assertion	De extra ziektekosten door roken moet je zelf betalen *You should pay for your own extra public health costs caused by smoking*
27	Formal assertion	Iemand is zelf aansprakelijk voor de extra ziektekosten die voortkomen uit roken[28] *A person is responsible for his own extra public health costs caused by smoking*
28	Conversational assertion	De extra ziektekosten door het drinken van alcohol moet je zelf betalen *You should pay for your own extra public health costs caused by drinking alcoholic beverages*
28	Formal assertion	Iemand is zelf aansprakelijk voor de extra ziektekosten die voortkomen uit het drinken van alcohol[29] *A person is responsible for his own extra public health costs caused by drinking alcoholic beverages*
29	Conversational assertion	De extra ziektekosten door nalatigheid in het verkeer moet je zelf betalen *You should pay for your own extra public health cost caused by careless driving*
29	Formal assertion	Iemand is zelf aansprakelijk voor de extra ziektekosten die voortkomen uit nalatigheid in het verkeer[30] *A person is responsible for his own extra public health costs caused by careless driving*

[28] Taken from original (formal) Dutch wording (Bernts 1991)
[29] Taken from original (formal) Dutch wording (Bernts 1991)
[30] Taken from original (formal) Dutch wording (Bernts 1991)

30	Conversational assertion	De extra ziektekosten door sporten moet je zelf betalen *You should pay for your own extra costs of public health caused by sports*
30	Formal assertion	Iemand is zelf aansprakelijk voor de extra ziektekosten die voortkomen uit sporten[31] *A person is responsible for his own extra public health costs caused by sports*
31	Conversational assertion	Ouderen zouden meer ziektekostenpremie moeten betalen dan jongeren *The elderly should pay more health insurance premium than youngsters*
31	Formal assertion	Ouderen (...) zouden meer moeten bijdragen aan de ziektekostenverzekering dan jongeren[32] *Elderly should contribute more to health insurance than youngsters*
32	Conversational assertion	Mensen die heel gezond leven door bijvoorbeeld te letten op hun eten, zouden minder moeten betalen aan de ziektekostenverzekering.[33] *People who live very healthy by for instance taking notice of their food should pay less for the health insurance*
32	Formal assertion	Mensen met een gezonde levensstijl, bijvoorbeeld met aandacht voor hun voeding, zouden minder moeten bijdragen aan de ziektekostenverzekering. *People with a healthy lifestyle, for instance with attention for their food, should contribute less to the health insurance*

[31] Taken from original (formal) Dutch wording (Bernts 1991)
[32] Taken from original (formal) Dutch wording (Elchardus et al., undated)
[33] Taken from original (conversational) Dutch wording (Elchardus et al., undated)

SECTION VII Health contacts choice questions		
33	(All versions)	**Introduction statement:** De volgende vragen gaan over uw contacten met de professionele gezondheidszorg. *The next question will be about your contacts in professional health care*
33	IFDN/IFEN **-Formal Q** **-Implicit alt.** **-**Difficult/Easy	Hoeveel jaar geleden is uw meest recente bezoek aan uw huisarts op het spreekuur? *How many years ago was your most recent visit to the G.P. on office hours?*
33	ICDN/ICEN **-Conv. Q** **-Implicit alt.** **-**Difficult /Easy	Hoeveel jaar geleden bent u voor het laatst bij de dokter op het spreekuur geweest? *How many years ago did you go to the doctor on office hours?*
33	FFEN/FFDN **-Formal Q** **-'Formal' alt** -Easy /Difficult	Heeft uw meest recente bezoek aan uw huisarts op het spreekuur korter dan een jaar geleden, tussen de een en twee jaar geleden, of langer dan twee jaar geleden plaatsgevonden? *Did your most recent visit to you G.P. on office hours take place shorter than a year ago, between one and two years ago or longer than two years ago?*
33	FCEN/FCDN **-Conv. Q** **-Formal alt** -Easy /Difficult	Was de laatste keer dat u bij de dokter op het spreekuur bent geweest, korter dan een jaar geleden, tussen een en twee jaar geleden, of langer dan twee jaar geleden? *Was the last time you went to the doctor on office hours shorter than a year ago, between one and two years ago or longer than two years ago?*
34	IFDN/IFEN **-Formal Q** **-Implicit alt.** -Difficult/Easy	Hoeveel maanden geleden is uw meest recente bezoek aan uw tandarts? *How many months ago was your most recent visit to your dentist?*
34	ICDN/ICEN **-Conv. Q** **-Implicit alt.** -Difficult /Easy	Hoeveel maanden geleden bent u het laatst bij de tandarts geweest? *How many months ago did you go to the dentist the last time?*
34	FFEN/FFDN **-Formal Q** **-'Formal' alt** -Easy /Difficult	Heeft u het meest recent korter dan drie maanden geleden, drie tot zes maanden geleden, of langer dan zes maanden geleden uw tandarts bezocht? *Did you visit your dentist most recently shorter than three months ago, thee to six months ago or longer than six months ago?*
34	FCEN/FCDN **-Conv. Q** **-Formal alt** -Easy /Difficult	Bent u voor het laatst korter dan drie maanden geleden, drie tot zes maanden geleden, of langer dan zes maanden bij de tandarts geweest? *Did you go to your dentist for the last shorter than three months ago, thee to six months ago or longer than six months ago?*

SECTION VIII Body measures choice questions		
35	(All versions)	**Introduction statement:** Nu volgen enkele achtergrondvragen *Now some background questions*
35	IFEN **-Formal Q** -Implicit alt.	Wat is uw lichaamlengte?[34] *What is your body length?*
35	ICEN **-Conv. Q** -Implicit alt.	Hoe lang bent u?[35] *How long are you?*
36	IFEN **-Formal Q** -Implicit alt.	Wat is uw gewicht in kilo's? *What is your body weight in kilograms?*
36	ICEN **-Conv. Q** -Implicit alt.	Hoeveel weegt u? *How much do you weigh?*
37	FFEN **-Formal question,** **-Formal alt.**	Staat u positief, negatief of neutraal tegenover uw gewicht? *Do you think of your weight in a positive, negative or neutral way?*
37	CFEN **-Formal question,** **-Conv. alt.**	Beschouwt u zichzelf als tevreden of ontevreden over uw gewicht of maakt het u niet uit? *Do you consider yourself as satisfied or unsatisfied with respect to your weight or does it not matter to you?*
37	CCEN **-Conv. Q** **-Conv. alt.**	Bent u tevreden of ontevreden over uw gewicht of maakt het u niet uit? *Are you satisfied or unsatisfied about your weight or does it not matter to you?*
38	FCEN **-Conv. Q.** **-Formal alt.**	Vindt u zichzelf te licht, te zwaar of vindt u uw gewicht niet te licht en niet te zwaar? *Do you find yourself too light, too heavy or do you consider your weight as not too light and not too heavy?*
38	CCEN **-Conv. Q.** **-Conv. alt.**	Vindt u zichzelf te mager of te dik of vindt u uw gewicht wel goed? *Do you find yourself too skinny, too fat or do you think your weight is good?*
38	CFEN **-Formal Q.** **-Conv. alt.**	Beschouwt u zichzelf te mager of te dik of beschouwt u uw gewicht als goed? *Do you consider yourself too skinny, too fat or do you think your weight is good?*
38	FFEN **-Formal Q.** **-Formal alt.**	Beschouwt u zichzelf te licht, te zwaar of beschouwt u uw gewicht als niet te licht en niet te zwaar? *Do you consider yourself too light, too heavy or do you consider your weight as not too light and not too heavy?*

[34] Taken from original (formal) Dutch wording (VBO, TNO 2001)
[35] Taken from CBS health survey (Van den Berg and Van der Wulp, 2003, see http://www.cbs.nl/nl/publicaties/artikelen/maatschappij/gezondheid/revisie-pols-1999.pdf)

SECTION IX Background questions		
39	IFEN -**Formal Q** -Implicit alt.	Verricht u op dit moment betaalde beroepsarbeid?[36] *Are you currently employed?*
39	ICEN **-Conv. Q** -Implicit alt.	Heeft u op dit moment betaald werk?[37] *Do you have a paid job at this moment?*
40	ICEA **-Conv. Q** -Implcit alt. -**Ambiguous**	Wat is de hoogste opleiding die u heeft afgemaakt?[38] *What is the highest level of education that you completed?*
40	ICEN **-Conv. Q** -Implcit alt. -**Non-ambiguous**	Wat is de hoogste opleiding waarvan u het diploma heeft? *What is the highest level of education that you have a diploma for?*
40	IFEA -**Formal Q** -Implicit alt **-Ambiguous**	Kunt u het schooltype noemen van uw hoogst genoten opleiding die u heeft afgerond? *Can you tell me the type of school of your highest level of education that you completed?*
40	IFEN -**Formal Q** -Implicit alt **-Non-ambiguous**	Kunt u het schooltype noemen van uw hoogst genoten opleiding die u met een diploma heeft afgesloten? *Can you tell me the type of school of your highest level of education that you completed with a diploma?*
40	Implicit response alternatives	-geen opleiding afgemaakt of lagere school (basisonderwijs) *-no education or elementary school* -lager beroepsonderwijs (bijv. LEAO, LTS) leerlingenstelsel, kort MBO *-lower vocational education* -middelbaar algemeen- of beroepsonderwijs (bijv. MAVO/mulo, MEAO, MTS) *- vocational education* -voortgezet algemeen onderwijs (bijv. HAVO, HBS, VWO, atheneum, gymnasium) *-pre-university education* -hoger beroeps- en universitair onderwijs (bijv. HEAO, HTS, SA, PA, MO-A, MO-B) *-higher vocational education or university*

[36] Taken from original (formal) Dutch wording (Familie enquête Nederlandse Bevolking, Nijmegen 1998: http://www.niwi.knaw.nl/nl/maatschappijwetenschappen/steinmetzarchief/dddi/docs/p1583.pdf/p1583.pdf)

[37] Taken from original (conversational) Dutch wording (Normvraagstellingen VMO: http://www.marktonderzoekassociatie.nl/info1.html)

[38] Taken from original (conversational) Dutch wording (VBO, TNO 2001)

41	IFEN -**Formal Q**	Wat is uw geboortejaar?[39] *What is your year of birth?*
41	ICEN -**Conv.Q**	Hoe oud bent u?[40] *How old are you?*
42	IFEN -**Formal Q** -Implicit alt.	Uit hoeveel personen bestaat, inclusief uzelf, op dit moment, uw huishouden? *Including yourself, how many persons does your household consist of at this moment?*
42	ICEN -**Conv. Q** -Implicit alt.	Hoeveel mensen, met uzelf erbij, wonen nu bij u in huis? *How many people, with yourself included, live at your house now?*
43	IFEA **-Formal Q** -Implicit alt. **-Ambiguous**	Bent u of is iemand in uw huishouden in het bezit van een personenauto? *Do you, or anyone in your household own a car?*
43	ICEA -**Conv. Q** -Implicit alt. **-Ambiguous**	Heeft u, of iemand bij u thuis een auto*?* *Do you, or anyone at your home have a car?*
43	IFEN -**Formal Q** -Implicit alt. **-Non-ambiguous**	Alle gemotoriseerde voertuigen met kenteken meegerekend, bent u of is iemand in uw huishouden in het bezit van een auto of motor? *Including all licensed vehicles, do you or anyone in your household own a car or motorcycle?*
44	Not manipulated final question	Rookt u? Do you smoke?

[39] Taken from original (formal) Dutch wording (Dutch ESS pilot, 2002)
[40] Taken from original (conversational) Dutch wording (Eurobarometer: http://europa.eu.int/comm/public_opinion/archives/eb/eb58/eb58_netherlands.pdf)

Appendix 7-2 Design of the questionnaires

Questionnaire 1		**Questionnaire 2**		**Questionnaire 3**		**Questionnaire 4**		**Questionnaire 5**		**Questionnaire 6**	
I	CCEA *Conversational, ambiguous question with conversational alternatives*	I	CFEA *Formal ambiguous question with conversational alternatives*	I	CCEN *Conversational question with conversational alternatives*	I	CFEN *Formal question with conversational alternatives*	III	IFDN *Formal difficult question with implicit alternatives*	III	ICEN *Conversational question with implicit alternatives*
II	FFEN *Formal question with formal alternatives*	VI	FCEN *Conversational question with formal alternatives*	VI	CCEN *Conversational question with conversational alternatives*	II	CFEN *Formal question with conversational alternatives*	IV	CCEN *Conversational question with conversational alternatives*	IV	FCEN *Conversational question with 5 formal alternatives*
III	IFEN *Formal question with implicit alternatives*	III	IFEA Ambiguous formal question with implicit alternatives ICDN *Difficults conversational question with implicit alternatives*	III	ICEA *Ambiguous conversational question with implicit alternatives* IFDN *Formal difficult question with implicit alternatives* IFDA *Formal difficult & ambiguous question with implicit alternatives*	III	IFEN *Conversational question with implicit alternatives*	II	FCEN *Conversational question with formal alternatives*	II	CCEN *Conversational question with conversational alternatives*

Questionnaire 1		**Questionnaire 2**		**Questionnaire 3**		**Questionnaire 4**		**Questionnaire 5**		**Questionnaire 6**	
IV	CFEN *Formal question with conversational alternatives*	II	CCEN *Conversational question with conversational alternatives*	II	FCEN *Conversational question with formal alternatives*	IV	FCEN *Conversational question with 5 formal alternatives*	V	ICEN Conversational question with implicit alternatives ICDN *Difficult Conversational question with implicit alternatives*	V	IFEN *Formal question with implicit alternatives* IFDN *Formal difficult question with implicit alternatives*
V	IFEN Easy Conversational question with implicit alternatives ICDN *Difficult Conversational question with implicit alternatives*	V	CFEN *Formal question with conversational alternatives* FFDN *Formal difficult question with formal alternatives* IFDN *Formal difficult question with implicit alternatives*	V	FFEN *Formal question with formal alternatives* FFDN *Formal difficult question with formal alternatives* IFDN *Formal difficult question with implicit alternatives*	V	IFEN *Formal question with implicit alternatives* IFDN *Formal difficult question with implicit alternatives*	VI	FCEN *Conversational question with formal alternatives*	VI	CFEN *Formal question with conversational alternatives*
VI	CCEN *Conversational question with conversational alternatives*	IV	FFEN *Formal question with 5 formal alternatives*	IV	CFEN *Formal question with conversational alternatives*	VI	FFEN *Formal question with formal alternatives*	I	CFEN *Formal question with conversational alternatives*	I	CCEN *Conversational question with conversational alternatives*

Questionnaire 1		**Questionnaire 2**		**Questionnaire 3**		**Questionnaire 4**		**Questionnaire 5**		**Questionnaire 6**	
VII	FCEN *Conversational question with formal alternatives*	VII	FFEN *Formal question with formal alternatives*	VII	IFDN *Formal difficult question with implicit alternatives*	VII	ICDN *Difficult conversational question with implicit alternatives*	VII	FCEN *Conversational question with formal alternatives*	VII	ICDN *Difficult conversational question with implicit alternatives*
VIII	IFEN *Formal question with implicit alternatives* CFEN *Formal question with conversational alternatives*	VIII	IFEN *Formal question with implicit alternatives* FFEN *Formal question with formal alternatives*	VIII	ICEN *Conversational question with implicit alternatives* CCEN *Conversational question with conversational alternatives* FCEN *Conversational question with formal alternatives*	VIII	ICEN *Conversational question with implicit alternatives* CCEN *Conversational question with conversational alternatives* FCEN *Conversational question with formal alternatives*	VIII	ICEN *Conversational question with implicit alternatives* CFEN *Formal question with conversational alternatives*	VIII	IFEN *Formal question with implicit alternatives* FFEN *Formal question with formal alternatives*
IX	IFEN Formal question with implicit alternatives IFEA *Ambiguous formal question with implicit alternatives*	IX	IFEN *Formal question with implicit alternatives*	IX	ICEN *Conversational question with implicit alternatives* ICEA *Ambiguous conversational question with implicit alternatives*	IX	ICEN *Conversational question with implicit alternatives* IFEA *Ambiguous formal question with implicit alternatives*	IX	ICEN *Conversational question with implicit alternatives*	IX	IFEN *Formal question with implicit alternatives* IFEA *Ambiguous formal question with implicit alternatives*

Samenvatting

In dit proefschrift wordt het interactionele proces van vragen en antwoorden in survey interviews onderzocht. Dit proces vindt plaats in zogenaamde vraag-antwoord sequenties (V-A sequenties). Een V-A sequentie bestaat uit alle uitingen die het stellen en beantwoorden van een vraag betreffen. De V-A sequentie begint op het moment dat de interviewer een vraag uit de vragenlijst stelt en eindigt bij de volgende vraag.

Een 'paradigmatische' V-A sequentie (Schaeffer and Maynard, 1996) is vanuit het oogpunt van de onderzoeker ideaal. In zo'n V-A sequentie stelt de interviewer de vraag precies zoals verwoord in de vragenlijst en de respondent geeft een antwoord dat eenduidig door de interviewer ingevuld kan worden. Wanneer de V-A sequentie niet paradigmatisch verloopt, kunnen er fouten in de meting optreden. Een respondent kan bijvoorbeeld een antwoord geven dat niet past bij de in de vragenlijst geformuleerde antwoordalternatieven (mismatch antwoorden). Vervolgens kan de interviewer een bepaalde antwoordcategorie aan de respondent suggereren. In dat geval is het uiteindelijke antwoord van de respondent door de interviewer beïnvloed. De kwaliteit van de gegevens verkregen door middel van het interview kunnen dan op negatieve wijze beïnvloed worden.

Afwijkingen van de paradigmatische V-A sequentie bieden duidelijke indicatoren van problematische processen bij het beantwoorden van vragen. Onderzoek van het type problemen dat optreedt bij deze verbale afwijkingen kan inzicht bieden in de oorzaken en gevolgen van zulke problemen en de gevolgen voor de kwaliteit van de verkregen gegevens.

Uiteraard is de aanwezigheid van een groot aantal paradigmatische V-A sequenties geen garantie voor het ontbreken van meetfouten. Respondenten kunnen sociaal-wenselijk antwoorden of vragen verkeerd begrijpen zonder dat er een afwijking van de paradigmatische V-A sequentie plaatsvindt. Er kan dan niets worden afgeleid uit analyses van het verbale gedrag in V-A sequenties.

Theoretische uitgangspunten bij problematisch interviewer en respondentgedrag

In dit proefschrift worden vier onderzoeksvragen beantwoord: (1) "Welk type problemen in de interactie in survey interviews kunnen vanuit een theoretisch oogpunt verwacht worden?"; (2) "Wat is de meest geschikte methode om interactionele problemen in survey interviews te achterhalen?"; (3) "Welke problematische afwijkingen van de paradigmatische V-A sequentie komen het meeste voor, worden ze met name door de interviewer of door de respondent geproduceerd en hoe staan deze afwijkingen in relatie tot ander gedrag in de V-A sequentie?"; (4) "Welke theoretische verklaringen kunnen voor het voorkomen van problematische afwijkingen in V-A sequenties gevonden worden?".

In hoofdstuk 2 wordt, in antwoord op de eerste onderzoeksvraag, een overzicht van conversationele en cognitieve theorieën gegeven die relevant zijn ten aanzien van interactie in survey interviews. De cognitieve verwerking bij het beantwoorden van een survey vraag, is door middel van het vier stappen model van Tourangeau et al. (2000) beschreven. Deze cognitieve stappen kunnen via conversationele principes van invloed zijn op de interactie.

De eerste stap, het begrijpen van de vraag, kan bijvoorbeeld problemen veroorzaken voor respondenten waardoor de betekenis van de vraag niet duidelijk is. Op welke manier problemen met het begrijpen van de vraag door respondenten worden gecommuniceerd en hoe interviewers hier vervolgens mee omgaan hangt af van conversationele principes. Respondenten zullen minder expliciet om toelichting van de vraag verzoeken, naarmate zij meer geneigd zijn gezichtsverlies te vermijden (beleefd zijn en vermijden de ander te beledigen) of een 'satisficing' strategie volgen (minder moeite investeren in hun taak), een lage taak involvement hebben (minder motivatie om het interview serieus te benaderen) en ervaring hebben met 'wat u er zelf onder verstaat'-reacties van de interviewer (dat is een gestandaardiseerde uitleg van de betekenis van de vraag). Op dezelfde manier zullen interviewers expliciete toelichting van de vraag vermijden, naarmate zij geneigd zijn gezichtsverlies te vermijden, zichzelf binden aan standaardiseringsregels, of moeite hebben met het herkennen van de oorzaak van een begripsprobleem. Wanneer interviewers standaardiseringsregels niet volgen, kunnen zij suggestief doorvragen of zelfs voor de respondent beslissen wat diens antwoord zou moeten zijn.

Stap 2 en 3, het ophalen van relevante informatie en het vormen van een oordeel over de geschiktheid van deze informatie, kunnen waarneembaar zijn in de interactie wanneer respondenten een optelstrategie volgen. Tijdens het verbaliseren van zo'n optelling kunnen interviewers reageren nog voordat de respondent met een definitief antwoord is gekomen. Deze reactie kan een suggestie tot een antwoord bevatten, wat natuurlijk problemen geeft voor de kwaliteit van het verkregen antwoord. Bovendien kunnen respondenten cues geven, zoals aarzelingen, hun onzekerheid over het antwoord verbaal communiceren, of een antwoord dat niet overeenstemt met de antwoordalternatieven uit de vragenlijst geven (een 'mismatch' antwoord). Wederom geldt dat hoe meer interviewers gezichtsverlies vermijden, des te meer zij het doorvragen naar precies geformuleerde antwoorden vermijden en de antwoorden van de respondenten zelf invullen.

Stap 4, het formuleren van het antwoord, kan de interactie beïnvloeden volgens een conversationeel principe dat 'voorkeur voor overeenstemming' heet. Respondenten hebben de neiging hun eerste antwoord instemmend te formuleren, of zij beginnen met aarzelingen voordat ze hun niet-instemmende antwoord geven. Wanneer interviewers zo'n eerste reactie te vlug accepteren als antwoord, is het mogelijk dat respondenten deze eerste antwoorden nooit herstellen tot het bedoelde niet-instemmende antwoord. Tot slot kunnen respondenten het survey interview als

een alledaags gesprek zien. Deze zienswijze kan ervoor zorgen dat zij hun antwoorden uitgebreid toelichten en mismatch antwoorden geven.

Samenvattend, de theoretische modellen die zijn gepresenteerd in hoofdstuk 2 laten zien dat de interactie tussen interviewer en respondent een verscheidenheid aan vormen kan aannemen. Zowel conversationele als cognitieve principes kunnen verantwoordelijk zijn voor een groot aantal verschillende problemen, zoals verzoeken om toelichting, mismatch antwoorden, het onjuist oplezen van vragen, suggestief gedrag, etc. Zulke problemen kunnen op hun beurt nieuw problematisch gedrag veroorzaken.

Methoden om interactionele problemen in survey interviews te achterhalen
De tweede onderzoeksvraag, hoe het verloop van de interactie op een systematische manier geanalyseerd kan worden om de relaties tussen gedragingen te onderzoeken, komt in hoofdstuk 3 aan de orde. Om V-A sequenties op een kwantitatieve manier te analyseren, is het nodig het verbale gedrag van interviewer en respondent te coderen (behavior coding). Volgens deze procedure worden aan de hand van een codeerschema systematisch codes toegekend aan gedragingen in een V-A sequentie. De codes kunnen een pure feitelijke beschrijving van het soort gedrag omvatten, zoals ‘interviewer stelt vraag’, maar kunnen ook een evaluatieve component bevatten (bijvoorbeeld ‘interviewer stelt vraag op een suggestieve manier’).

In hoofdstuk 3 wordt ook een overzicht van verschillende methoden van behavior coding gegeven. Uit dit overzicht van de in totaal 48 gevonden codeersystemen blijkt dat er diverse beslissingen genomen dienen te worden ten aanzien van de te volgen procedures en strategieën. Een eerste beslissing betreft het selectieve karakter van het codeerschema. Er kan gekozen worden voor volledige codering (‘full coding’) wat inhoudt dat alle uitingen worden gecodeerd, of voor selectieve codering (‘selective coding’); alleen een selectie van, in het licht van bepaalde onderzoeksvragen belangrijk geachte gedragingen, wordt gecodeerd.

Een tweede beslissing, die ten dele afhangt van de eerste, is de eenheid van analyse. Codering kan plaats vinden op het niveau van de uiting, het ‘beurt-uitwisselingsniveau’ (dat is elke beurt, ongeacht het aantal specifieke uitingen) of de hele V-A sequentie. Vervolgens moeten er beslissingen genomen worden over het behoud van sequentiële informatie, praktische procedures (‘live’ coderen, van geluidsbestanden, of met behulp van transcripten) en het type codeurs (interviewers, de onderzoeker of getrainde codeurs). Voor welke codeerstrategie gekozen wordt, heeft gevolgen voor de analysemogelijkheden van de gecodeerde data. Wanneer snelle informatie beoogd wordt (bijvoorbeeld in de pretest fase van onderzoek of bij het monitoren van dataverzameling) kan al voldoende informatie uit frequentieanalyses verkregen worden. Bij dergelijke analyses wordt alleen nagegaan hoe vaak een bepaalde code voorkomt, al dan niet in relatie tot bepaalde vraag-, interviewer- of respondentkenmerken. Geschikte codeerschema’s zijn dan beperkt tot

'selective coding' (met minder dan 15 codes). Codering kan ook na afloop van het hoofdonderzoek plaatsvinden, in het kader van de evaluatie van het dataverzamelingsproces. Hoewel bij dergelijk onderzoek snelle resultaten minder belangrijk zullen zijn, zal een gedetailleerde analyse van de oorzaken van problematische gedragingen ook dan over het algemeen niet nodig zijn. In dat geval kan een selectief codeerschema met een iets groter aantal codes (bijvoorbeeld ongeveer 20) geschikt zijn.

In het geval van exploratieve analyses van de interactie is gedetailleerde informatie gewenst en lijkt 'full coding' met behoud van sequentiële informatie het meest geschikt. Voor de praktische toepassing van zulke codeerschema's is software beschikbaar, zoals het Sequence Viewer programma (zie Dijkstra 2002).

Omdat onze onderzoeksvragen gericht zijn op een volledige beschrijving van de interactie en ook verwijzen naar de volgorde van voorkomen van gedragingen, is het nodig een 'full coding' schema met behoud van sequentiële informatie te gebruiken. Dit is, vanwege de sequentiële informatie, de meest informatieve, maar ook de meest arbeidsintensieve manier van coderen. Het gekozen codeerschema wordt beschreven in hoofdstuk 4. Dit schema voldoet aan alle criteria van bruikbaarheid en de gewenste hoeveelheid detail in de codes. Met dit multivariate codeerschema (Dijkstra 1999) worden gedragingen gecodeerd op een aantal verschillende variabelen. Elke variabele beschrijft een bepaald aspect van de uiting. De combinatie van waarden levert een code-string op die een betekenisvolle beschrijving van de uiting geeft. Het multivariate karakter van het schema maakt het ook makkelijker te wisselen tussen grove analyses (gebruikmakend van een deel van de codeervariabelen) en gedetailleerde analyses (gebruikmakend van de meeste of alle variabelen). Het schema is bovendien redelijk gemakkelijk te gebruiken door codeurs. Er wordt geïllustreerd hoe het codeerschema gebruikt kan worden om vrijwel elk voorkomend gedrag dat relevant is voor het verloop van de interactie, te coderen.

Samenvattend, de codeer procedures en strategieën, en de codes zelf, hangen in hoge mate af van het doel of de focus van het onderzoek. Omdat het onderzoek in dit proefschrift gericht is op het beschrijven van de interactie tussen interviewer en respondent, hebben we besloten om een zeer gedetailleerd, volledig codeerschema te gebruiken, met de uiting als analyse-eenheid en met behoud van sequentiële informatie.

Oorzaken van problematische afwijkingen van de paradigmatische V-A sequentie

De derde onderzoeksvraag, over de meest frequent voorkomende problematische afwijkingen en hun oorzaken is het onderwerp van hoofdstuk 5. Hiervoor werden telefonische interviews van een onderzoek naar gedrag en houding ten aanzien van televisiekijken en reclame gebruikt. De exploratieve analyses van de getranscribeerde en gecodeerde interviews toonden aan dat in bijna 50% van de V-A sequenties problematische afwijkingen voorkwamen.

De eerste problematische afwijking in een V-A sequentie werd over het algemeen door de respondent geproduceerd. Dit was met name een mismatch antwoord. Als er een mismatch antwoord wordt gegeven, zijn interviewers genoodzaakt door te vragen tot zij een adequaat antwoord krijgen. Deze noodzaak leidt er vaak toe dat interviewers suggestief doorvragen of ander problematisch gedrag vertonen. Blijkbaar is het zeer belangrijk interviewers te trainen in het adequaat omgaan met mismatch antwoorden.

Een nog betere strategie om de kwaliteit van de gegevens te verbeteren is het optreden van mismatch antwoorden te voorkomen. Vraagformulering en het type antwoordalternatieven dat gebruikt wordt spelen hierbij een belangrijke rol. Over het algemeen leveren antwoordalternatieven die niet goed zijn afgestemd op de vraag een groot aantal mismatch antwoorden op. Bij een vraag die als ja-nee vraag is geformuleerd horen 'ja' en 'nee' als alternatieven te zijn opgenomen en geen specificaties van 'ja' of 'nee' (bijvoorbeeld 'ja, altijd', 'ja, soms', etc.). Antwoordalternatieven met een vier of vijfpunts Likert-type schaal ('zeer mee eens'-'mee eens'-'mee oneens'-'zeer mee oneens') leveren ook een groot aantal mismatch antwoorden op. Bovendien bleek er een grotere kans op mismatch antwoorden wanneer interviewers niet eenduidig geïnstrueerd worden hoe zij, in een serie vragen met dezelfde antwoordalternatieven, de antwoordalternatieven voor iedere vraag dienen te herhalen. Ja-nee vragen bleken daarentegen een laag aantal mismatch antwoorden op te leveren, indien de antwoordalternatieven bestonden uit 'ja' en 'nee' en eventueel 'weet niet' of 'weigert te antwoorden'.

Oorzaken van andere problematische afwijkingen bleken ook gerelateerd te zijn aan vraag- en respondentkenmerken. Bijvoorbeeld als algemene vragen volgen op specifieke vragen slaan interviewers de algemene vraag vaak over of vragen onvoldoende door. Daarnaast bleken oudere, laag opgeleide en vrouwelijke respondenten meer problematische afwijkingen te produceren dan jongere, hoogopgeleide en mannelijke respondenten.

Samenvattend is het antwoord op de derde onderzoeksvraag dat mismatch antwoorden verreweg de meest voorkomende problematische afwijkingen zijn en bovendien een belangrijke oorzaak van problematisch interviewer gedrag. Vraagformulering, maar vooral het type antwoordalternatieven, bleken naast respondentkenmerken de belangrijkste oorzaak van mismatch antwoorden.

Theoretische verklaringen voor het optreden van mismatch antwoorden

Het veelvuldig voorkomen van mismatch antwoorden en het feit dat zij ook de belangrijkste oorzaak van problematisch interviewer gedrag zijn, motiveerde ons verder onderzoek te doen naar mogelijke oorzaken van deze problematische afwijking, om onze vierde onderzoeksvraag te kunnen beantwoorden.

In hoofdstuk 6 worden drie oorzaken voor het optreden van mismatch antwoorden beschreven. Ten eerste kan er een conversationeel probleem optreden. We veronderstellen dat dit de belangrijkste oorzaak van het optreden van mismatch antwoorden is. Respondenten hebben over het algemeen geen flauwe notie over wat er van hen verwacht wordt in een survey interview. Zij verwarren het gestandaardiseerde interview met een alledaags gesprek en formuleren hun antwoord daarom zoals zij gewend zijn te doen in gewone gesprekken. Bijvoorbeeld, wanneer respondenten gevraagd wordt hoeveel dagen per week zij televisie kijken, dan kunnen zij denken dat het geoorloofd is een antwoord als 'de meeste dagen' te geven, in plaats van een precies gedefinieerd aantal dagen. Zo'n antwoord is echter niet codeerbaar voor de interviewer omdat het niet overeenkomt met de beschikbare antwoordalternatieven. Deze mismatch antwoorden worden conversationele mismatch antwoorden genoemd.

Ten tweede kan ambiguïteit in de betekenis van de vraag taakonzekerheid veroorzaken. Hoewel respondenten alle relevante informatie beschikbaar hebben met betrekking tot de vraag, zien zij zich zelf geconfronteerd met het probleem om hun specifieke (complexe) situatie te vertalen in een van de antwoordalternatieven die hen worden aangeboden. Als gevolg van deze onzekerheid kunnen respondenten verbale overwegingen geven en daarmee de kans op een mismatch antwoord vergroten. Deze mismatch antwoorden worden taak mismatch antwoorden genoemd.

Ten derde kan er een cognitief probleem optreden, wanneer de gevraagde informatie niet direct beschikbaar is in het geheugen. In dat geval kunnen respondenten enerzijds beginnen hardop te denken en ook overwegingen bij het antwoord geven en daarmee de kans op een mismatch antwoord vergroten. Anderzijds kunnen zij, zonder hard op denken meteen een mismatch antwoord (bijvoorbeeld in de vorm van een schatting) geven ten einde het probleem te vermijden. Deze mismatch antwoorden worden cognitieve mismatch antwoorden genoemd.

Op grond van bovenstaande drie oorzaken van mismatch antwoorden zijn een aantal hypothesen opgesteld over de effecten van specifieke formuleringen van vragen en antwoordalternatieven op de kans van het optreden van mismatch antwoorden. Deze hypothesen werden getoetst in een non-experimenteel en een experimenteel onderzoek.

Het non-experimentele onderzoek (hoofdstuk 6) betrof interviews van het Nederlandse vooronderzoek van de European Social Survey (ESS). De vragenlijst van deze CAPI-interviews ('computer aided personal interviewing') bevatte 268 verschillende vragen, waarbinnen verschillende vraagcategorieën konden worden onderscheiden om de hypothesen non-experimenteel te kunnen toetsen. In het experimentele onderzoek (hoofdstuk 7) werden vraagformuleringen gebruikt die afkomstig waren van daadwerkelijk bestaande onderzoeken over gezondheid. Er werden meerdere versies van een zelfde vraag opgesteld om de effecten van vraagformulering en type alternatieven te kunnen vergelijken.

Conversationele mismatch antwoorden: Vraagformulering
Verondersteld wordt dat conversationeel geformuleerde vragen (geformuleerd zoals gebruikelijk is in gewone gesprekken) respondenten het idee geven dat een conversationele manier van antwoorden, minder exact en precies, acceptabel is. Formele vragen daarentegen, attenderen respondenten erop dat een exact en precies antwoord vereist is. De hypothese is daarom dat conversationele vragen meer mismatch antwoorden genereren dan formele vragen.

Deze hypothese kon worden bevestigd in de non-experimentele analyse van de ESS gegevens, voor vragen waarbij geen toonkaarten gebruikt werden. De vragen in dit onderzoek die beoordeeld werden als 'conversationeel', leverden meer mismatch antwoorden op dan vragen die als 'formeel' beoordeeld werden. In het experimentele onderzoek kon de hypothese alleen bevestigd worden voor enkele opiniestellingen en enkele achtergrondvragen, maar in het algemeen werden effecten van het conversationele karakter van de vraag niet gevonden. Voor een aantal vragen bleken de resultaten de hypothese zelfs tegen te spreken.

Onze manipulaties van het conversationele karakter van de vragen waren gebaseerd op het gebruik van alledaagse woorden en wij trachtten tegelijkertijd realistische survey vragen te formuleren. We vermoeden dat deze manipulaties niet extreem genoeg waren. Veel van de conversationele vragen kunnen in feite nog steeds worden beschouwd als vragen waarvan het onwaarschijnlijk is dat ze, zo geformuleerd, gesteld worden in gewone gesprekken. In enkele gevallen werden dus in werkelijkheid 'formele' vragen met 'nog formelere' vragen vergeleken. Manipulaties op basis van complete zinnen (dus ook de conversationele grammaticale structuur en kans op voorkomen in alledaagse conversaties in ogenschouw nemende) zou tot duidelijker resultaten hebben moeten leiden.

Conversationele mismatch antwoorden: type alternatieven
Behalve het op een formele manier formuleren van vragen om conversationele mismatch antwoorden te voorkomen, kan ook juist gebruik gemaakt worden van conversationele antwoordalternatieven. Op die manier kan voorkomen worden dat respondenten vragen moeten beantwoorden op een onalledaagse manier. Antwoordalternatieven die in alledaagse gesprekken veel gebruikt worden kunnen de kans op mismatch antwoorden verminderen. Formele woorden worden juist minder vaak gebruikt in alledaagse gesprekken. Onze hypothese was daarom dat vragen met conversationele alternatieven minder mismatch antwoorden opleveren dan vragen met formele alternatieven.

Deze hypothese kon in zowel het non-experimentele als het experimentele onderzoek bevestigd worden. In het experimentele onderzoek werden de sterkste effecten gevonden voor stellingvragen. Het blijkt dat respondenten stellingvragen

gewoonlijk als ja-nee vragen behandelen (dat wil zeggen met 'ja' of 'nee' beantwoorden), terwijl 'ja' en 'nee' ook typisch conversationele woorden zijn.

Survey vragen kunnen worden gesteld met een expliciet lijstje alternatieven of met impliciete alternatieven. Vragen van het laatste type hebben een open vraagstelling waarbij de range van alternatieven geïmpliceerd wordt (bijvoorbeeld het aantal uren of minuten, een percentage, een aantal dagen etc.). De alternatieven worden echter niet expliciet door de interviewer genoemd. In alledaagse gesprekken is het erg ongebruikelijk om antwoordalternatieven op te sommen. Er werd daarom verondersteld dat impliciete alternatieven conversationeler zijn dan een expliciete lijst met alternatieven. Volgens de hypothese zouden vragen met impliciete alternatieven minder mismatch antwoorden opleveren dan vragen met expliciete alternatieven.

Deze hypothese kon noch in het non-experimentele, noch in het experimentele onderzoek bevestigd worden.

Taak mismatch antwoorden

Vragen met ambigue concepten kunnen taakonzekerheid veroorzaken waardoor respondenten moeite kunnen hebben met het kiezen van alternatieven. De hypothese dat vragen waarin ambigue concepten niet gespecificeerd worden meer mismatch antwoorden opleveren dan vragen waarin dit wel gebeurt, kon bevestigd worden in het non-experimentele onderzoek. Vragen afkomstig uit de ESS vragenlijst die beoordeeld werden als 'ambigue' leverden meer mismatch antwoorden op dan vragen die beoordeeld werden als 'niet ambigue'. Het verschil in het percentage mismatch antwoorden was echter klein. Bovendien kon aan de hand van de verbale interacties geen aanwijzing gevonden worden dat de mismatch antwoorden inderdaad veroorzaakt werden door taakonzekerheid. In het experimentele onderzoek kon de hypothese voor slechts één van de zeven gemanipuleerde vragen bevestigd worden. Voor één vraag werden zelfs tegengestelde resultaten gevonden. De niet-ambigue versie van deze vraag bevatte zoveel specificaties (over wat als een 'auto' beschouwd kan worden), dat meer onzekerheid bij respondenten gecreëerd werd dan wanneer geen specificaties gegeven werden. De manipulatie van ambiguïteit had dus andere dan de bedoelde effecten. Bovendien heeft de lagere frequentie van optreden van taak-mismatch antwoorden als gevolg dat het moeilijker is vraagformuleringseffecten te vinden.

Cognitieve mismatch antwoorden

Respondenten kunnen moeite hebben met het ophalen van informatie die nodig is om de vraag te beantwoorden. Deze moeite kan door middel van verbale overwegingen gecommuniceerd worden, waarbij een grotere kans ontstaat op het optreden van mismatch antwoorden. Onze hypothese luidde dat vragen waarvoor informatie nodig is die niet onmiddellijk in het geheugen beschikbaar is (moeilijke vragen) tot meer mismatch antwoorden zullen leiden dan vragen waarvoor relatief gezien minder

cognitieve moeite hoeft te worden gedaan (makkelijke vragen). Deze hypothese kon worden bevestigd met de ESS gegevens. Vragen uit dit onderzoek welke werden beoordeeld als 'moeilijk' leverden meer mismatch antwoorden op dan vragen die werden beoordeeld als 'makkelijk'.

In het experimentele onderzoek kon de hypothese echter niet bevestigd worden. Voor de vergelijking van 'makkelijke' en 'moeilijke' versies van vragen werden slechts twee verschillende vragen gemanipuleerd. Hoewel uit verzoeken om toelichting en weet-niet antwoorden van respondenten bleek dat de 'moeilijke' vragen inderdaad als moeilijker werden beschouwd dan de 'makkelijke' vragen, verschilden de vragen niet in het aantal mismatch antwoorden dat zij opleverden. Het bleek niet eenvoudig om vraagversies te creëren die betrekking hebben op dezelfde informatie, maar verschillen met betrekking tot moeilijkheid.

Aanbevelingen

De resultaten beschreven in dit proefschrift toonden aan dat het eerste probleem in een V-A sequentie meestal door de respondent wordt veroorzaakt en dat dit voornamelijk mismatch antwoorden betreft. Als gevolg van deze mismatch antwoorden vertonen ook interviewers problematisch gedrag (in de vorm van suggestief doorvragen of antwoorden veronderstellen zonder ze te verifiëren), wat negatieve gevolgen kan hebben voor de kwaliteit van de verkregen antwoorden. Bovendien, zelfs wanneer interviewers adequaat weten te reageren op mismatch antwoorden om adequate antwoorden te krijgen, heeft deze reactie een verlengde interactie tot gevolg en daarmee hogere kosten van survey interviews. Om de kans op mismatch antwoorden te verminderen of de negatieve gevolgen ervan te beperken kunnen de volgende aanbevelingen worden gedaan:

- Gebruik antwoordalternatieven die goed bij de vraag passen. Wanneer een vraag als ja-nee vraag is geformuleerd, moeten de antwoordalternatieven uitsluitend bestaan uit 'ja', 'nee' en, indien van toepassing, 'weet niet' en 'weigert te antwoorden'.
- Gebruik antwoordalternatieven die aangepast zijn aan de conversationele manier van antwoorden. Zulke alternatieven bestaan uit woorden die het meest gebruikt worden in alledaagse gesprekken.
- Voor stellingvragen kunnen het beste 'ja' en 'nee' als antwoordalternatieven gebruikt worden. Stellingen kunnen ook met aangepaste formulering gebruikt worden, zoals "in welke mate vindt u dat…". Er is een grote kans op mismatch antwoorden wanneer stellingvragen in een pure stellingformulering worden geformuleerd samen met een vijf-puntsschaal. Respondenten zien zulke vragen als ja-nee vragen, wat inhoudt dat zij typisch antwoorden met 'ja' en 'nee'. Zulke antwoorden passen niet bij een vijf-punts Likert schaal.
- Werk in een vragenlijst niet vaker dan noodzakelijk met afwisselend verschillende rijtjes antwoordalternatieven. Rijtjes die enigszins op elkaar lijken

(bijvoorbeeld de rijtjes ‘eens’-‘neutraal’-‘oneens’ en ‘waar’-‘mogelijk waar’-‘niet waar’) maar bijvoorbeeld variëren in aantal antwoordopties (bijvoorbeeld wel of geen middelste alternatief) zijn verwarrend voor respondenten, wat tot meer mismatch antwoorden leidt. Er kunnen ook ‘buffer vragen’ gebruikt worden, met een compleet ander antwoordformaat (bij voorkeur open vragen met impliciete antwoordalternatieven) tussen blokken vragen met verschillende antwoordalternatieven.

- Wanneer vragen met formele alternatieven onvermijdelijk zijn, gebruik ze dan niet vóór vragen met conversationele alternatieven. Het blijkt namelijk dat respondenten na een aantal vragen wel wennen aan formele alternatieven, en ze dan blijven gebruiken bij latere vragen, ook als dat vragen met conversationele alternatieven zijn.
- Gebruik in face-to-face interviews toonkaarten met antwoordalternatieven.
- Gebruik een formele vraagformulering, om respondenten aan het formele karakter van het interview te herinneren, wat hen zal stimuleren meer precies geformuleerde antwoorden te geven.
- Leer interviewers om mismatch antwoorden te herkennen, en train ze in het gebruik van adequate reacties (zoals: ‘herhaal de antwoordalternatieven en biedt altijd meer dan één alternatief aan’).

Author index

Subject index